PROPOSING AN ACHIEVEMENT SIMULATION METHODOLOGY TO ALLOW THE ESTIMATION OF INDIVIDUAL IN CLINICAL TESTING CONTEXT

Enio Galinkin JELIHOVSCHI[1]
Cristiano Mauro Assis GOMES[2]

- ABSTRACT: Information from the population should be directly transposed to the individual level only under strict conditions of stationarity and homogeneity. In general, psychological phenomena are neither stationary nor homogeneous. Furthermore, the individual parameters must be estimated. The usual techniques of estimating the individual are rarely attainable. This occurs because in order to get valid estimates of individual parameters, many occasions of the same test must be performed. In this article, we propose an approach (*simerg* algorithm) to estimate the individual, so that a proper clinical testing for the individual in the context of testing setting that require the respondents' performance becomes viable. We present the rationale of *simerg* and apply it to estimate the verbal comprehension ability of three individuals, which performed, in 90 different occasions, the 60 items of this ability. In this application, we show how *simerg* uses the initial empirical performance of these persons and simulates their whole performance. We evaluate the validity of *simerg* using different measures to predict the empirical performance. We conclude that *simerg* produces valid estimates. Further studies in a variety of populations and tests that require respondents' performance are needed to enable a stronger conclusion about the validity of *simerg*.

- KEYWORDS: Rasch models, simulation, ergodicity, individual testing.

[1] Universidade Estadual de Santa Cruz - UESC, Departamento de Exatas e Ciências Tecnológicas, CEP: 45662-900, Ilhéus, BA, Brazil. E-mail:*eniojelihovs@gmail.com*

[2] Universidade Federal de Minas Gerais - UFMG, Departamento de Psicologia, CEP: 31270-901, Belo Horizonte, MG, Brazil. E-mail:*cristianomaurogomes@gmail.com*

1 Introduction

A psychological testing of individuals is understood as the practice of making inferences about a particular person by using psychological tests. The traditional practice of psychological testing has long been the way of making inferences using as references the population estimations of psychological tests (URBINA, 2014). This approach is so ingrained in the contextualization of the way psychologists perform testing, that it has been considered to be the only way to do psychological testing. In this approach each individual in a population is tested just once, so that only population parameters are estimated. Individuals results are considered to be just a replication of the same random variable. This kind of individual testing which takes as reference population estimated parameters is known as norm-referenced test interpretation. In Urbina (2014) at page 83 own words: “Norms are, by far, the most widely used frame of reference for interpreting test scores”. In fact, this way of practice underlines the logic of construction, inference and interpretation of test scores; what is really estimated is the population, not the individual. This psychological framework uses the estimates of population parameters as if they were also the individual´s. Therefore, the individual parameters do not have to be estimated even though he or she is the main target to be measured by the testing, especially in clinical testing contexts, where the clinician aims to produce inference about certain characteristics of a particular individual, regarding the clinical questions that sustain and justify the motive of the psychological testing. Citing Urbina (2014) again: “The score itself is used to place the test taker’s performance within a preexisting distribution of scores or data obtained from the performance of a suitable comparison group”. At this point, we must point it out, that we are not reducing the psychological assessment to the exclusive use of tests nor to the practice of comparing the individual to normative criteria. We are perfectly aware that the psychological evaluation is a broad process of inferences that uses information from psychological tests, as well other sources of information, such as clinical interview, history information, observation and qualitative data. Nevertheless, a new movement has been taking place for the last decades in psychological testing of individuals.

Ergodic theory has been extensively used by Molenaar (2004, 2007) and other authors e.g. (FISHER, 2018) to show that estimation of a person’s trait cannot accurately been made by making use of averages from population results. If we want to validly estimate a particular person’s ability, or to estimate the test difficulty for this person, this should be done by repeatedly measuring this person. According to the ergodic theorems, information from the population should be directly transposed to the individual level only under strict conditions of stationarity and homogeneity. In general, psychological phenomena are neither stationary nor homogeneous. Furthermore, the individual parameters must be estimated. This condition causes serious implications in clinical testing contexts, leading Gomes et al. (2018), through a detailed presentation of the ergodic theorems, to conclude that the routine psychological testing of the individual is not valid. According to Gomes et al. (2018), performing a valid psychological testing of the individual implies that the individual parameters must be estimated, which means applying the same test a great number of occasions. Repeating many times the application of a test involving the respondents' performance is a procedure which is rarely attainable, e.g. time series factor analysis (MOLENAAR, 2007) as well as nonlinear state-space models (MOLENAAR, 2010) and dynamic Rasch model (RIJN, 2010) require 100 or more testing occasions, that in practice no researcher will be able to repeat a testing procedure so many times. This points to the importance of a valid simulation of those testing results, i.e. starting from a few repetitions of the testing procedure we may be able to simulate a large number of testing occasions.

That is the goal of this paper, i.e. the presentation and the initial application of an algorithm, created by the first author of this article, named *simerg*, which seeks to make it feasible a valid psychological testing of the individual. It means, a practice which enables the clinician to make inferences about the individual through the own person's estimation. Through *simerg*, 3 to 6 occasions of testing will be sufficient to estimate the individual. To reach this objective, we simulate a series that represents the respondent's performance in the same psychological test applied to the same person many times, starting with only 3 and 6 occasions of the empirical respondent's performance.

In this article, we present the rationale of the algorithm *simerg* and show how an implementation of it in *R* (R CORE TEAM, 2018) is applied to estimate the verbal comprehension ability of three individuals, which performed, in 90 different occasions, the 60 items of a three marker tests of this ability. In this application, we show how *simerg* uses the initial empirical performance of these persons and simulates their whole performance. We evaluate the validity of simerg through its accuracy to predict the empirical performance of the individuals.

The small number of individuals, the empirical results, shows how difficult is get people to repeat a testing for 90 times. Nevertheless, we find that this small number still makes it possible to get meaningful results. The numbers 3 and 6 are just values enough to make it feasible the practice of repeated testing, and to run the algorithm.

2 Algorithm *simerg*: Rationale

The Rasch model for dichotomous data has the following mathematical form:

$$Pr(\mathbf{X}_n i = 1) = \frac{\mathrm{e}^{\theta_n - \delta_i}}{1 + \mathrm{e}^{\theta_n - \delta_i}} \tag{1}$$

where θ_n is the ability of person n, δ_i is the difficulty of item i, $\mathbf{X}_n i = x$ for $x \in \{0, 1\}$ and 0 denotes an incorrect answer and 1 a correct answer.

The two parameters θ_n and δ_i are usually estimated. Since our interest is based upon repeating the same test we will focus on θ, that is, we will try to find how the person´s ability grows as he or she answers the same test. Even though the results of testing repeatedly the same person are not independent, the estimated parameters of Rasch models do capture the learning process as the individual keeps working on the same test. As Mair (2007) say: “When operating with longitudinal models, the main research question is whether an individual’s test performance changes over time. The most intuitive way would be to look at the shift in ability θ across time points.” As we have been pointing out in this article, we propose a new method in order to make the estimation of individual parameters viable in the context of clinical testing. This method consists in a simulation of individual performance from just a small number, 3 to 6, of empirical testing occasions. To that purpose we wrote an algorithm, called *simerg*, that is, simulation in a non ergodic psychometric process. *simerg* simulates what would be the individual score for 90 or more occasions by using only the information of a few empirical testing occasions. Hence, *simerg* saves the psychologist from applying the test up to 100 times in order to estimate the parameters he or she is interested to study, concerning the individual.

The algorithm *simerg* uses four parameters as follows:

- v - data.frame containing the empirical data. Each line of v is one occasion of a given testing and each column is a respective item of the testing. In this case, each cell of v represents the individual’s performance in a particular item in a specific occasion. All cells have the values 0 or 1; the value 0 indicates a wrong answer while the value 1 a correct answer.
- k - number of starting rows from the empirical dataset to be used as start for the simulation process. It is defined by the user.
- n - number of interactions when creating groups of ability parameters in order to get the simulated data.frame with the desired size (number of rows).
- *ca* - the calibration parameter, an integer ranging from 2 to k+3 and defined by the user.

The pseudo code of the algorithm *simerg* is:

```
1. initialize with user specifications values for v, n, k, ca
2. use the first k rows of v, call this as vf; this vf will become the simulated set
   of data
3. estimate the vector of ability parameters $\theta$ from vf; the vector contains k
   values; call this vector th2
4. add the corresponding rows to a expanded vf according to the 3 greatest values
   of the ordered $\theta$ ´s
5. now start the procedure to complete vf until the desired number of rows
   (occasions)
```

6. **for** *i in 1:n*

7. initially the vector of ability parameters θ are estimated by using vf with k+3 rows, 3 more rows will be added every cycle, the same for the vector th2 but th2 starts with only k values

8. estimate the ability parameters θ ´s from vf, which has k + 3 rows at the start of the loop, and so, a vector of size k + 3 of θ are estimated.

9. simulate a dataset using Rasch model and the vector of the estimated θ values

10. write the estimated vector of ability parameters θ in ascending order; if x is the size of the vector, choose the 3 values in the positions $x - ca$, $x - (ca - 1)$ and $x - (ca - 2)$

11. add to vf the 3 rows according to the chosen values θ ´s

12. add those 3 chosen values of θ to the vector th2

13. **end the loop**

14. the estimation is repeated once more to adjust for the number of occasions and another time just to get the last values of theta so that the right number of simulated occasions and parameters θ´s are obtained.

The parameters n and k are related through the formula: first let the number of simulated occasions be called nso so that $nso = 3(n+2)+k$ or $n = \frac{nso-k}{3} - 2$. In certain cases, the number of desired simulated occasions will be exactly produced by *simerg*, which is the case of the example. However, in other cases, *simerg* will only produce a close but not exactly the number desired. The parameter ca represents the learning rate of the individual and is connected to how optimistic the *simerg* function becomes. If ca equals 2 *simerg* becomes very optimistic and goes up to the maximum number correct answers very fast. As ca increases *simerg* becomes less optimistic, producing a learning rate like a logistic curve, with a strong learning in the beginning that becomes more and more weak, reaching a plateau, which indicates a stabilization of the ability development.

The algorithm *simerg* was implemented in *R* language (R CORE TEAM, 2018) into a function also called *simerg* where the estimation of the ability parameter $\theta$and simulation of new occasions were done by using the functions RM and *sim.rasch* from *eRm* package (MAIR, 2007). If the individual is learning from occasion to occasion, this will appear as a constant shift in the respective values of the θ, in spite of the function RM assuming independence among the testing occasions. The result of function *simerg* is a list of two elements. The first element is a dataframe containing the simulated series of testing occasions; the second is a vector containing the respective abilities per row generated for every entry row of the first element of the data.frame which corresponds to every test simulated plus the first k rows of the empirical data. The function *simerg* tries to simulate in the same way as people learn when submitted to many testing occasions. This means that the number of correct answers and the ability parameter θ increase as the testing goes on until they arrive to a plateau. After that, this number remains constant.This process may happen in many different measures, speed, quality or quantity, nevertheless it always goes to one direction.

The function *simerg* and goodness of fit calculations were written in *R* (R CORE TEAM, 2018) using the Tinn-R graphical user interface (FARIA, 2008).

3 Validity of the *simerg* Function

The usual method to check the validity of a Rasch model is the likelihood ratio test which is carried out by dividing the maximum value of the likelihood function under the restricted model by the maximum value of the likelihood function under the global model, see Bickel (2001), sections 4.9, and 6.3.1. In general those calculations are almost impossible to be carried out, which reduces the use of that test only to models under the assumption of independence among the many realizations of the observed values which, in our case, are the testing occasions, see Agresti (2002), page 24. In this case, the likelihood function of the complete sample becomes the product of that function calculated for just one testing occasion as many times as the total number occasions. Unfortunately, in our case, this cannot be done since the temporal dependence of the testing repetitions, which is performed by the same participant. Therefore, we do not apply the traditional criteria for inspecting the validity of the *simerg* algorithm, which is embedded into the unidimensional and dichotomous Rasch model, aiming to inspect if the unidimensional model is valid to estimate the individual. We do not use the likelihood ratio test statistic - LR, see Andersen (1973), as well the infit and outfit values. We nonetheless use a descriptive goodness of fit criterion that informs about the quality of the prediction of the applied model into the algorithm *simerg*. Information about the predictive capacity of the model have been used as an alternative criterion, regarding the validity of the model, as recommended by Mair (2008). Therefore, we evaluate the validity of the function *simerg*, concerning the unidimensional and dichotomous Rasch model applied in the estimation of the individual, just inspecting the following predictive criteria: (1) accuracy, (2) sensitivity, (3) specificity. All of them bring information about the quality of the prediction of the algorithm *simerg* to predict the answers of the individual. To do that, we compare the empirical data performance of the individual to her or his simulated performance produced by the *simerg* algorithm. First let us define some useful notation.

- If the person passed an item and *simerg* correctly predicts that the person passed this item, it is called a true positive prediction (TP), or type 1-1.
- If the person failed an item and *simerg* correctly predicts that the person failed this item, it is called a true negative prediction (TN), or type 0-0.
- If the person failed an item and *simerg* erroneously predicts that the person passed this item, it is called a false positive prediction (FP), or type 0-1.

- If the person passed an item and simerg erroneously predicts that the person failed this item, it is called a false negative prediction (FN), or type 1-0.
- The confusion matrix from package *caret* (KUHN, 2018) shows the accuracy, sensitivity, specificity between two series (empirical versus simulated), where a correct answer is marked 1, whereas a wrong answer is marked 0.

The goodness of fit ways to see how well the *simerg* simulated series estimate the empirical series are the following:

Definition 3.1. Accuracy shows the proportion of correctly predicted observations, that is
$accurracy = \frac{TP+TN}{TP+TN+FP+FN}$, where TP+TN+FP+FN represent all empirical observations.

Definition 3.2. Sensitivity shows the percentage of the positive (correct answers) empirical observations correctly predicted (TP), that is,
$sensitivity = \frac{TP}{TP+FN}$, where TP+FN represent all empirical positive observations.

Definition 3.3. Specificity shows the percentage of the negative (wrong answers) empirical observations correctly predicted (TN), that is,
$specificity = \frac{TN}{TN+FP}$, where TN+FP represent all empirical negative observations.

Definition 3.4. Let **f** and **g** be two curves where both are plotted on the same coordinate axes where the horizontal axis points are $(x_1, x_2, \ldots, x_n)$ so the mean sum of squares (MMS) between **f** and **g** is:

$$\frac{1}{n}\sum_{i=1}^{n}(f_i - g_i)^2 \tag{2}$$

In our case, f is formed by the θ´s which represent the empirical ability growth curve and g is formed by the θ´s which represent the *simerg*'s simulated growth curve. Those θ´s are represented in logits and in the x axis the points represents the 90 occasions.

Accuracy is the most important measurement for inspecting the validity of simerg ,because it measures the ratio of all *simerg* correct predictions versus all empirical observations. Sensitivity and specificity, on the other hand, act like a magnifying glass within the accuracy, they show which type of correct predictions are more abundant; either the type 1-1 or the type 0-0. They complement the accuracy in the sense that if the number of type 1-1 correct answers decreases and number of correct answers of type 0-0 increases on the same proportion, the accuracy remains unchanged.

It is up to the user to define which type of correct prediction is more important in his or her study; in ours, the empirical and simulated testing of this paper, the type 1-1 correct answer regarding sensitivity, is more important than the type 0-0, concerning specificity, because it reflects the natural movement of the learning process within the empirical which is going from 0 (wrong answer) to 1 (right answer). This movement produces the variation needed to estimate the parameters. We must have, at the end, much more correct answers than wrong answers, since otherwise the tests are too difficult for that individual as it was for one of the participants. This will be shown in the next sections.

4 Participants and empirical data

Three participants responded three tests (total of 60 items) that measure the same latent variable, that is, verbal comprehension, during 90 different occasions; it lasted three months and the participants had contact with the tests only at the moment of the administration. For each occasion of testing, if the person passed the item, it was scored as one for this item; otherwise, this item was scored as zero. For a complete description of the test as well the validity studies regarding them, see Gomes et al. (2014).

The participants of the study have different educational backgrounds, age and sex. The $participant_j$ is a 23 year old male student, who graduated in biological sciences, the $participant_d$ is a 60 year old female, a retired physical education teacher, and the $participant_q$ is 34 year old female, who has worked as a house maid and administrative secretary; she studied only up to 8th grade. This small numbers of participants shows how difficult it is to get participants in any testing situation when the number of occasions, where the participant have to redo the test, is large, as in our case.

The applied tests are briefly described:

- V_1: It consists of 24 items and the time limit for its execution is 6 minutes. Each item is composed of one reference word and five multiple-choice words. The respondent must identify the word which best approximates, in terms of meaning, to the reference word and mark it with an X.

- V_2: It consists of 18 items and the time limit for its execution is five minutes. Each item is composed of one reference word and five multiple-choice words. The respondent must identify the word which best approximates, in terms of meaning, to the reference word and mark it with an X.

- V_3: It consists of 18 items and the time limit for its execution is five minutes. Each item is composed of one word of reference and four multiple-choice words. The respondent must identify the word which best approximates, in terms of meaning, to the reference word and mark it with an X.

5 Application and strategies

The *R* command to run the function *simerg* is:

vs = simerg(v, n= , k= , ca=)

v is the dataframe containing the empirical data. As previously explained each line of v is one occasion of a given testing and each column is a respective item of the testing; *n* is the number of runs so that it will achieve the number of occasions desired; *k* is the number of initial rows (empirical occasions) to start the process of simulation.

When applying the simerg function to each of the three participants, we used the command *simerg(v, n, k, ca)* only changing the dataframe of the parameter v, that is, the 90 empirical occasions regarding the performance of each participant. The user must be warned that sometimes the function RM might break and thus forcing the *simerg* function to stop. He or she only have to run it again, it usually works.

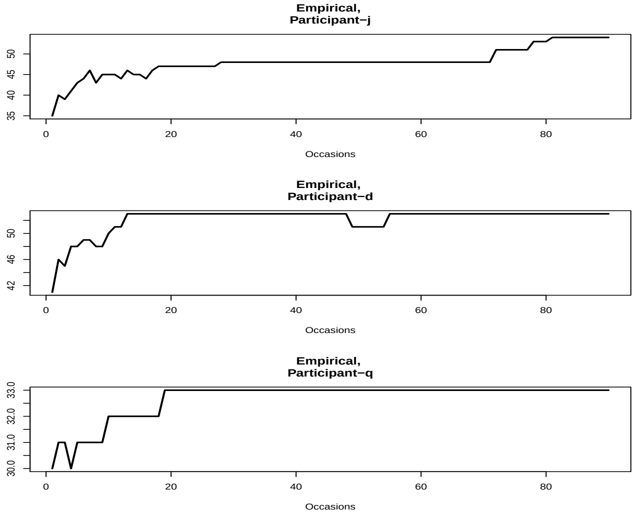

Figure 1 - Sum of responses for each testing occasion.

Before evaluating the validity of simerg, regarding its capacity to correctly predict the performance of the participants, the empirical performance of each one of these participants in the 90 testing occasions is shown in figure 1. This description is pertinent since it indicates the intra-individual variation present in the performance. It also implicitly shows whether the pool of 60 items was adequate for estimating the target construct, that is, verbal comprehension, for each participant. Figure 1 shows that $participant_d$ learned faster than $participant_j$. She reached 53 correct answers and stayed there for 72 occasions while $participant_j$ stayed 44 occasions on 48 correct answers and then went slowly to 54 correct answers staying there for 10 occasions. $Participant_q$, on the other hand, is another story. She reached only 33 correct answers, about half of total numbers of questions, and stayed there for 72 occasions. The questions were indeed too difficult for her. We may then conjecture that the 60 items seem to work better for *participant*$_j$, somewhat worse for *participant*$_d$ and badly for *participant*$_q$ in terms of their power to generate a large intra-individual variation.

Since the testing occasions of the *participant*$_q$ did not have enough variation in order to get its parameters estimated by function RM, *simerg* did not work for her. This situation occurs because the RM function uses a maximum likelihood estimation which assumes independent identically distributed variables. In our case, each variable is an occasion result. Since, as we explained above, that is not the case, a lot of variation is lost when the same person responds for 90 occasions. Therefore, the test must be well designed for the population in which it will be applied, so that some variation is left to be used in the estimation process.

Next we present the simerg simulation, using k = 3 and k = 6, regarding the performance of the *participant*$_d$ and the *participant*$_j$. As we stated previously, we are simulating the behaviour of simerg as k = 3 and comparing this to k = 6, because the parameter k sets the stage for the real testing. The smaller is the value of k, k = 3 being the minimum value for k, the easier to carry out the clinics of the individual. If the performance of the simulation with k = 3 is valid to estimate the individual, then it means that in practice he or she will have the participants to respond for only three occasions, so that the clinics of the person becomes viable in practice.

We evaluate the validity of simerg by inspecting its accuracy to predict the real performance of those individuals. The accuracy is the central information when inspecting the validity of simerg since this information shows the proportion of correctly predicted observations and it has been used as an alternative criterion for evaluating the validity of a model (MAIR, 2008). Moreover, the sensitivity and specificity are additional information, but neither central nor really important to infer whether simerg is valid or not. Assuming that the accuracy is central to infer the validity of simerg, we need to stress that there is no cutoff value defined by the literature which allows the researcher to infer whether a prediction model is valid to estimate the individual. This absence occurs in the fields of machine learning, data mining, psychometrics, and so on. Since we do not have a cutoff value, we do not know if a model which predicts 5% or 30% or 90% of the individual performance is valid for measuring the target construct that is being measured in this individual. However, this is a problem that can be solved. Despite of not having a cutoff value for our specific study, we know that the psychometrics literature is plenty of cutoff values for other studies. For example, psychometrics assumes that if the items of a test show a intra-class correlation of 0.70, then the researcher can infer that this test is reliable (URBINA, 2014). A more important example for our problem, psychometrics considers that if a test correlates at least 0.75 with a gold standard test, then the researcher can conclude that the first test measures the same construct as it is measured by the gold standard test (URBINA, 2014).

Table 1 - Performances for different values of *ca*

		$k = 3$				$k = 6$			
	ca	2	3	4	5	2	3	4	5
$Participant_j$	Accuracy:	0.80	0.75	0.75	0.64	0.82	0.82	0.77	0.76
	Sensitivity:	0.96	0.87	0.87	0.70	0.98	0.98	0.91	0.88
	Specificity:	0.14	0.28	0.26	0.40	0.16	0.16	0.22	0.25
	MSS:	11.83	4.36	6.40	9.64	15.42	19.90	6.49	6.95
	ca	2	3	4	5	2	3	4	5
$Participant_d$	Accuracy:	0.77	0.77	0.83	0.72	0.83	0.79	0.81	0.79
	Sensitivity:	0.85	0.85	0.92	0.78	0.91	0.87	0.89	0.86
	Specificity:	0.23	0.26	0.17	0.35	0.24	0.25	0.26	0.29
	MSS :	1.60	2.04	1.69	3.14	1.68	1.63	1.33	1.78

Let us look at the accuracy as a measure of approximation between the empirical and the simulation series. Accuracy measures the proportion of correctly predicted observations among all empirical observations. That means that accuracy plus the proportion of falsely predicted observations among all empirical observations equals one. If all predictions are correct then accuracy equals one if all predictions are false then accuracy equals zero. In that sense ,making an analogy to the coefficient of determination in linear regression, the accuracy accounts for the variation of the simulated model due to the empirical and 1 minus accuracy is the proportion of the variation in the simulated series which cannot be accounted by the empirical since they are falsely predicted. In linear regression, a coefficient of determination of 0.70 (70%) means a correlation of 0.84. In our case, however, we cannot make this formulation but the analogy leads us to infer that a accuracy of 0.70 is a very good cutoff value for a criterion of model validity.

Another way to display the results is by making graphical output from the θ values (ability parameters) generated by *simerg* and the θ's of the empirical values estimated by the RM function. Those graphs are called ability growth curve, because they show the growth of that ability parameter.

Table 1 shows the accuracy, sensitivity and specificity for k = 3 and k = 6 and different values of ca. It clearly shows that when *ca* increases from two to five the accuracy decreases, the sensitivity decreases but the specificity increases. This occurs because accuracy is connected to the correct predictions and predictions of type 1 - 1 decreases more than type 0 - 0 increases, that is sensitivity goes down and specificity goes up because, as *ca* increases, *simerg* becomes less optimistic and, therefore it's results has less ones and more zeroes.

Considering the cutoff value of 0.70 for accuracy, we infer that simerg is valid to simulate the verbal comprehension ability both for $partipant_d$ and $partipant_j$, for all values of ca and k except for ca = 5 and k = 3 in $partipant_j$. Moreover, from information from Table 1 we conclude that the accuracy for the conditions of ca = 2, ca = 3, and ca = 4, are very close to each other, for k = 3 and k = 6, except when ca = 5 and k = 3, where the accuracy decreases meaningfully and fall below

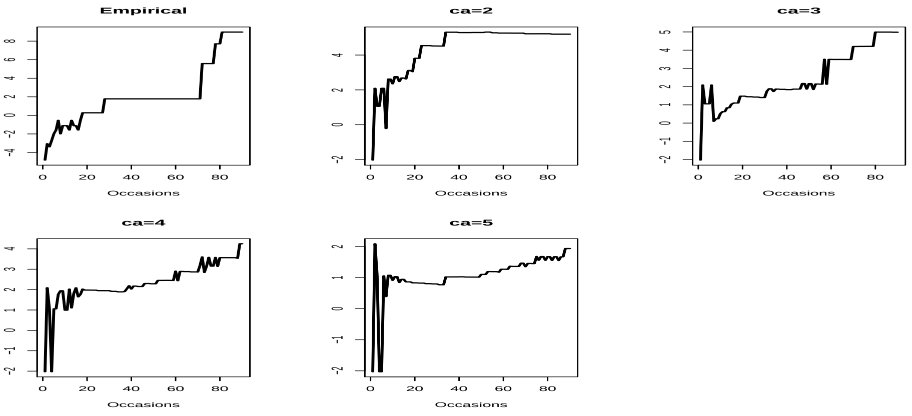

Figure 2 - $Participant_j$, k=3.

the cutoff value of 0.70 for the $partipant_j$.

Nonetheless, we should not focus only on Table 1. In order to evaluate the validity of simerg, we first must find out whether or not simerg is capable to simulate the real performance of the individual, in terms of its trajectory. If simerg simulates trajectories very far from the real performances, it will be difficult to sustain that simerg is valid. In this case, figures 2, 3, 4 and 5 are very valuable, since they show the real performance of the individuals, as well as the simulated performance on the conditions of k = 3, k = 6 and ca = 2, 3, 4, 5. Those figures explain the behaviour of both participants helping us, for instance, to observe that the empirical performance of both participants are similar to a logistic curve, with a strong learning at the beginning that becomes weaker, reaching a plateau, indicating a stabilization of the learning. They show very clearly how learning is taking place, where are the places of great variation and when actual learning occurs. Furthermore, the graphs show the occurrence of the *ceiling effect* in the simulation results.

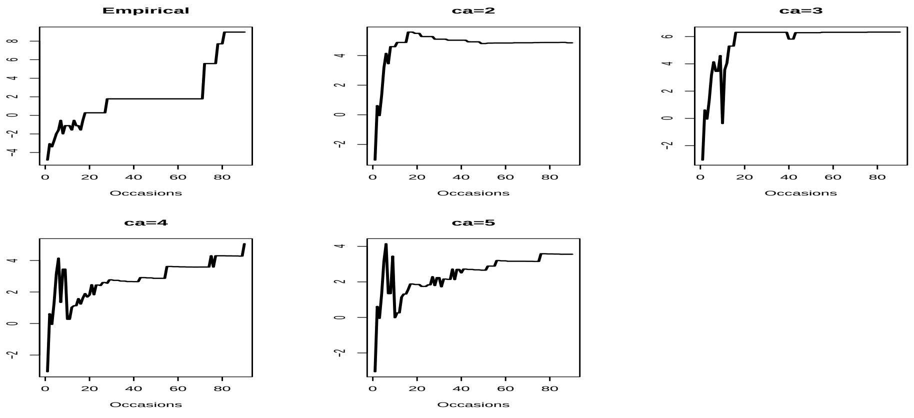

Figure 3 - $Participant_j$, k=6.

Before continuing, we should pay attention that the empirical performance of the participants do not achieve a sequence of 60 points - the maximum score, called *ceiling effect*, indicating that this effect is not a property of the true performance of the participants. For example, $partipant_j$ achieves 35 points in the first testing occasion, that is a fast progress in the beginning. In the seventh testing occasion, he shows a strong increment of 9 points in comparison to the first occasion, performing 46 points. After that, he continues to progress, but in a slower rate, stopping at the 81th testing occasion, when he performs 54 points, he never reaches 60 points. If a ceiling effect takes place then the graphs would never stop going up should more questions and repetitions be given. In this case, the graph shoots up to the top and stays there what will surely depart from the empirical behaviour. It is very important to note that the ceiling effect will depend on the values of the θ's as marked on the vertical axis. For example, for $participant_d$, k=6 and ca=3, the maximum value of θ is a little above 3 logits, therefore there is no ceiling effect even though the oscillation of the θ's at the top is accentuated. On the other hand, for ca=4 the maximum value of the θ's is 5 logits, and therefore the occurrence of the ceiling effect. For $participant_j$ figures 2 and 3 show that for ca = 2, 3 the simulations produce ceiling effect (the value of gets very close or equal to 5 logits) showing that they do not fit the true performance. The same, for ca = 2,3 and 4 can be deduced for $participant_d$ This implies that those simulated trajectories are inadequate to represent the true performance of the individuals, and are not valid. We should, therefore, be cautious when looking only on the results of table 1.

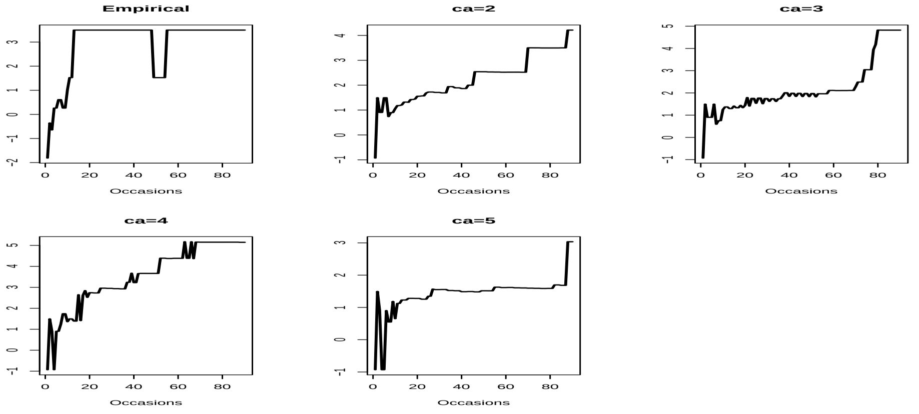

Figure 4 - *Participant$_d$*, k=3.

The simulation of *participant$_d$* for $k = 3$ and $ca = 4$ seem somewhat anomalous. We ran *simerg* many tinmes, the results remained close to each other. This may be a result from the way the three first occasions of *participant$_d$* empirical grow in number of ones (right answers), which, in connection to the function RM estimation procedure and the value of *ca*, produce that anomaly. It shows that this simulation is very optimistic that is, the ammount of 1´s goes up very fast, even faster than the simulation using $ca = 2$, towards 60 right answers, causing an early ceiling effect. This anomaly accounts for the results shown in table 1.

Participants j and d have very different backgrounds and age so that, as we have already commented, the test is not well suited to both, it seems to be better suited to *participant$_j$*. This clearly reflects on the results of the simulation. For example, for $k = 3$ *participant$_j$*'s simulation shows the ceiling effect up to $ca = 3$. For $ca = 4$ it looses the ceiling effect while still mantaining its main quality which is an ever increasing number of correct answers until reaching a plateau similar to what happens in practice. For $ca = 5$ it oscillates too much, what shows that for *participant$_j$* and $k = 3$, $ca = 4$ is the best choice. On the other hand, for *participant$_d$* and $k = 3$, only $ca = 5$ does not have the ceiling effect, nonetheless it oscillates too much. For $k = 6$, as it is expected, the simulation results are more stable as is shown in table . In this case, for *participant$_j$* both $ca = 4$ and $ca = 5$ do not show the ceiling effect and $ca = 5$ does oscillates but under acceptable limits. For *participant$_d$* $ca = 3$ and $ca = 5$ exhibits the best results.

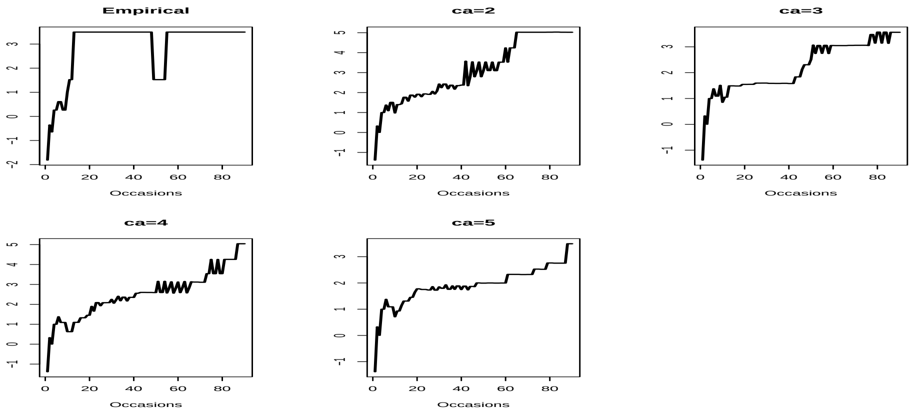

Figure 5 - *$Participant_d$*, k=6.

Table 1 also shows the MSS, which is a quantitative measure of the closeness between two curves and sheds light on what could be called the *participant's effect*. It is interesting that MSS´s are consistently smaller for *$participant_d$* than for *$participant_j$*. This shows that algorithm *simerg* is dependent on the participant, when simulating tests, to get simulated ability growth curves close to the empirical ability growth curves. We must take into account that *simerg* uses only the first 3 or 6 occasions of the 90 occasions of both participants and figure 1 shows that *$participant_d$* have a steeper beginning than *$participant_j$*, that is, at the beginning *$participant_d$* is more optimistic than *$participant_j$* and this helps the algorithm to get smaller ASS for her. For k = 3, since the simulated curves are too erratic, the results cannot be taken into account. For k = 6 the smallest results are those corresponding to ca = 4 and 5 except for *$participant_d$* ca = 5, but the difference between it and the ones for ca = 2 and 3 is small. This certainly agrees with the conclusions based on erratic behaviour or ceiling effect.

From the results above we may infer that it is up to the user to design the test according to the population who is going to respond to it, and get the right calibration of *simerg*.

Finally, we may conclude that simerg is valid for the values of k = 6 and ca = 5. The value of k = 6 enables a consistent simulation, since there are more empirical data to sustain it. The ca = 5 is adequate, since its use results in a series which is not too optimistic and therefore do not produce any ceiling effect, which is consistent with the empirical performance. The value of ca = 5 tend to produce more variability in terms of correct answers, but the initial value of k = 6 corrects this variability resulting in a good simulation. As a conclusion, we recommend the minimum value of k = 6 and ca value of 5, because they produce a stable series without ceiling effect.

Conclusion

In this paper we claimed that, if the clinician wants to produce a correct inference about the individual (FISHER, 2018; GOMES et al., 2018; MOLENAAR, 2004, 2007), it is mandatory that she or he estimates the individual. We also stated that the usual methods for estimating the individual demand a huge number of testing occasions, which led us to claim that the clinics of the individual is, nowadays, an almost impossible task. This happens mainly in the context of testing sets that demand the individual's performance. This situation led us to raise the challenging question: how can we then estimate the individual? To answer this question, we proposed an algorithm named *simerg*, with the intention to produce an initial attempt to make viable the clinics of the individual. The rationale of *simerg* shows that it was created to simulate the individual's performance in tests that measure a particular construct in the population.

We introduced the algorithm *simerg*, its parameters and how to use the function implemented in *R*. We applied simerg to two individuals who had performed a set of 60 items of verbal comprehension through a series of 90 occasions, expecting that simerg would be useful in simulating the real performance of these individuals. We analyzed the results of the simulation using suitable methods of exploratory data analysis - accuracy and graphical analysis of the simulated trajectories in comparison to the empirical trajectories - , since the dataset do not fit to be analyzed by confirmatory methods - likelihood ratio test statistic, infit and outfit. We found initial evidence that simerg seems to be a promising tool to make viable the clinics of the individual. The values of k = 6 and ca = 5 of simerg were valid for simulating the verbal comprehension of two participants of the study. This initial evidence is promising, since it points out that the clinician would estimate the individual through only six testing occasions, which seems to us something feasible and reasonable.

We wish to lay emphasis on the fact that our paper represents only the very beginning of a long way that must be treaded until the achievement of a solid and viable clinics of the individual. Until now, there is nothing in the literature proposing a methodology that make possible the clinics of the person, regarding testing sets that require individuals' performance. So that, one of the relevances of this article lays in the fact that it is the first one to propose a particular methodology to do that. However, there are many challenges laid down in next studies and we invite other researchers to use *simerg* and try their hand in formulating other approaches.

The first one is to enlarge the validity studies of *simerg*, considering the application of simerg in a variety of samples (e. g. adults, children, older people, educated people, uneducated people). The second one is to find out whether *simerg* reliably simulates the individual's performance in the context of the measurement of different abilities (e. g. reasoning, spatial, memory, planning, monitoring). The third one seeks to understand what is the relationship between the difficulty of the test and the learning capacity of the individual. For example, $participant_q$ showed very low intra-individual variation, she began with 30 correct answers and finished with only 33 correct answers. We do not know whether this low learning rate is the exception or the rule and, up to now, we do not know how to deal with this lack of variation. The low learning rate is a serious problem in the context of the clinics of the individual, since the intra-individual variation is the "heart" of the estimation of the person; without variation, there is no measurement. If the individual shows, at the moment, a low capacity to learn, he or she, probably, will not change the performance through the many testing occasions and will not produce variation.The fourth is a relevant effort of investigation regarding the understanding whether the first testing occasions are central or determinant to estimate the entire trajectory of the individual's learning. We observed that as the participants of our study learned with the items, their performance changed mainly and considerably at the beginning of the testing occasions. After the 20th occasion there were very little improvement on the performance of these individuals. Considering that the individuals' trajectory is a crucial property to estimate the individual, we need to understand whether what we have found represents a general property that can be considered when simulating the individual performance or it is only a singular characteristic of our participants. We acknowledge that the number of participants is small, nonetheless we find the results powerful enough to deserve attention.

We hope that this article will encourage other researchers to study how to make viable the clinics of the individual. We also hope that the same researchers find *simerg* useful in their research. *simerg* was written in *R* language which is open source and its code may be easily changed according to the user needs. We also provide *simerg* as well as all data and code used in this article, so that anyone who deem it interesting, may replicate the results and, therefore learn and understand how *simerg* works.

Acknowledgments

We thank the reviewers and editors for their very good comments and suggestions which helped in improving the paper. Cristiano Mauro Assis Gomes thanks CNPq-Brazil for its productivity fellowship.

JELIHOVSCHI, E. G.; GOMES, C. M. A. Proposta de uma metodologia de simulação de desempenho para a estimação do indivíduo no contexto da testagem clínica. *Rev. Bras. Biom.,* Lavras, v.37, n.4, p.493-511, 2019.

- *RESUMO: A informação da população pode ser transposta diretamente para o indivíduo somente dentro de condições estritas de estacionaridade e homogeneidade. Os fenômenos psicológicos, em geral, não são nem estacionários e nem homogêneos, mesmo assim os parâmetros individuais terão de ser estimados, porém, as técnicas de estimação são raramente possíveis. Isso ocorre, porque para se obter válidas estimativas dos parâmetros individuais, o mesmo teste terá de ser repetido em várias ocasiões com o mesmo indivíduo. Neste artigo, propomos uma nova bordagem (algoritmo* simerg*) para estimar o indivíduo no sentido de tornar viável o teste clínico, dentro dos contextos que exigem desempenho do indivíduo que responde aos testes. A base lógica do algoritmo é mostrada e depois aplicada para estimar a habilidade da compreensão verbal de três indivíduos que que fizeram, em 90 ocasiões, os 60 items desta habilidade. Nesta aplicação mostramos como* simerg *usa o desempenho inicial destas pessoas para simular o desempenho total. Também avaliamos a validade do* simerg*, usando algumas medidas, como preditora do desempenho empírico. Concluimos que* simerg *pode produzir estimativas válidas. Mais pesquisas com populações e testes de desempenho mais variados são necessários para que se possa chegar a conclusões mais robustas sobre a validade do* simerg*.*

- *PALAVRAS-CHAVE: Modelos Rasch, simulação, ergodicidade, testagem individual.*

References

AGRESTI, A. *Categorical data analysis.* 2.ed. New Jersey: Wiley-Interscience, 2002.

ANDERSEN, E. B. A goodness of fit test for the rasch model. *Psychometrika*, v.1, n.38, p.123–140, 1973.

BICKEL, P. J.; DOKSUM, K. A. *Mathematical statistics basic ideas and selected topics - Volume I.* 2.ed. New Jersey: Prentice-Hall, 2001.

FARIA, J. C. *Tinn-r gui/editor for R language and environment* statistical computing 2016. http://nbcgib.uesc.br/lec/software/editores/tinn-r/en

FISHER, A. J.; MEDAGLIA, J. D.; JERRONIMUS, B. Lack of group-to-individual generalizability is a threat to human subjects research. *PNAS*, v.115, n.27, p.1–10, 2018.

GOMES, C. M. A.; ARAUJO, J.; FERREIRA, M. G.; GOLINO, H. F. The validity of the cattel-horn-carroll model on the intraindividual approach. *Behavioral Development Bulletin*, v.19, n.4, p.22-30, 2014.

GOMES, C. M. A.; ARAUJO, J.; NASCIMENTO, E.; JELIHOVSCHI, E.G. Routine psychological testing of the individual is not valid. *Psychological Reports*, v.122, n.4, p.1476–1593, 2018.

KUHN, M. *caret: Classification and regression training*, 2018. https://CRAN.R-project.org/package=caret.

MAIR, P.; HATZINGER, R. Extended rasch modeling: The erm package for the application of irt models in R. *Journal of Statistical Software*, v.20, n.9, p.1–20, 2007.

MAIR, P.; REISE, S. P.; BENTLER, P. M. Irt goodness of fit using approaches from logistic regression. *Department of Statistics Papers*, UCLA, 2008. https://escholarship.org/uc.

MOLENAAR, P. C. M. A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, v.2, n.4, p.201–218, 2004.

MOLENAAR, P. C. M. On the implications of the classical ergodic theorems: analysis of developmental process has to focus on intraindividual variation. *Development Psychobiology*, v.50, n.1, p.60–69, 2007.

MOLENAAR, P. C. M.; RAM N. Dynamic modeling and optimal control of intra- individual variation: A computational paradigm for nonergodic psychological processes. In CHOW, S. M.; FERRER, E.; HSIEH, F.(Ed.). *Statistical methods for modeling human dynamics: An interdisciplinary dialogue*, p.13–37. New York: Routledge, 2010.

R CORE TEAM. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, 2018. https://www.R-project.org/.

URBINA, S. *Essentials of psychological testing.* New Jersey: John Wiley & Sons, 2014.

VAN RIJN, P.; DOLAN, C. V.; MOLENAAR, P. C. M. State space methods for item response modeling of multisubject time series. In MOLENAAR, P. C.; NEWELL, K. M. (Eds.). *Individual pathways of change: Statistical models for analyzing learning and development*, p.125–151. Washington DC: American Psychological Association., 2010.

Received on 16.03.2019.

Approved after revised on 29.10.2019.