Retrieving IRT parameters with half information

Marie Sacksick, Sebastián Ventura

Retrieving IRT parameters with half information

Marie Sacksick[*]
KDIS - University of Cordoba
CHArt - University Paris 8
marie.sacksick@domoscio.com

Sebastián Ventura
KDIS - University of Cordoba
sventura@uco.es

ABSTRACT

Item Response Theory can be used to estimate the degree of mastery of a concept by learners, to automatically assess their knowledge. The models stemming from this theory are tuned to be adapted to the questions used to assess mastery. The correct estimation of the parameters is key to be able to have a correct estimation of the mastery. However, this estimation can be skewed by missing data, noise on the model, or a lack of data.

The question we ask here in this paper is how much data, created by a given number of students answering to a given number of questions is necessary to retrieve reliable coefficients of the questions, when the database at disposal have missing data. To do so we use simulated data. There are two case studies with different levels of data emptiness: one is the baseline and has complete information, the other has only half information.

We find that even though IRT models seem robust against missing values, it is not possible to use the thresholds of the literature obtained with a full database.

Keywords

item response theory, parameter estimation, missing values

1. INTRODUCTION

Item Response Theory (IRT) models are used in psychometrics to evaluate the value of a "latent trait", the value of a descriptor that cannot be assessed directly. IRT offers a framework to be able to measure this unreachable feature. These models are widely used in education to evaluate the degree of understanding and of mastery of a piece of knowledge. In the educational context, that latent trait is called "ability".

As the "ability" cannot be assessed directly, it is necessary to know from which amount of data it is possible for the model to give good estimations. Some studies have been conducted, such as (Chuah, Drasgow, & Luecht, 2006) and (Şahin & Anıl, 2016), to highlight a threshold of data amount under which the results cannot be seen as reliable. To our knowledge, no study has been conducted with a database where the students only answered some of the questions, and not all the database. This situation is very likely to happen, for example when the learner did not have enough time to answer all the questions.

This study uses simulated data, which therefore respects exactly the IRT model. We are investigating whether the IRT algorithm is able to retrieve the simulated questions coefficients. Data is simulated, and cleaned; we run an IRT algorithm thanks to the software R; and finally the theoretical and experimental parameters are compared and the quality of the estimation is estimated through various indicators.

2. RELATED WORK

2.1 What is IRT?

The IRT builds a probabilistic model which hypotheses a relationship between characteristics of the questions and the mastery of the topic by the student. This model has two sets of parameters: the latent trait dedicated to the representation of the student, and some dedicated to the representation of the questions. In this study, we only focus on unidimensional models, and the latent trait can be represented by a unique parameter, usually noted θ.

Here, given a student S_j and given a question Q_i and working with the unidimensional 2-Parameter Logistic model, the probability of success of the student for that question can be written:

$$P(S_j, Q_i) = \frac{e^{(a_i(\theta_j - b_i))}}{1 + e^{(a_i(\theta_j - b_i))}} \tag{1}$$

With $[a_i, b_i]$ being the parameters of the question Q_i. a_i is called the discrimination and it is positive. b_i is called the difficulty; it can either be positive or negative, and 0 represents the mean difficulty.

The inputs of the item response theory models are the answers of the students to the questions. The likelihood of the responses patterns given the probability of success explained in eq. (1) is maximized so as to deduce the most likely questions coefficients and students abilities. Compared to other evaluation theories, one of the advantages of the IRT models is their ability to deal with missing values. There are many methods to estimate the parameters, including Bayesian or non-Bayesian, so we will not list them here.

2.2 Previous work

[*] PhD Student hosted by the society Domoscio

We would like the reader to keep in mind that, depending on the algorithm used to estimate the parameters, the parameters will not be the same, so as their precision; Gao and Chen (2005) give an example of such a situation. Given the evolution in the methods of parameters estimation, we choose not to use studies older than 2000. This drastically reduces the number of studies trying to evaluate the fit of the estimation of the parameters.

RMSD is used in various studies as indicator: (Svetina et al., 2013) and (Yavuz & Hambleton, 2017) use it on simulated data to compare theoretical and experimental values of item parameters; (Svetina et al., 2013) also uses it on person ability. It is also used in (Şahin & Anıl, 2016) on the item parameters where the baseline is the parameters obtained when all the data are used; the estimation using part of the database are said good when $RMSD \leq 0.33$. (Wyse & Babcock, 2016) uses it to have a view on items parameters variations.
Correlation is used as an indicator by (Şahin & Anıl, 2016) like RMSD; the results are said good when $r \geq 0.70$.
Biais can be also used, like in (Svetina et al., 2013).

These studies never paid attention to the influence of missing values, while in education, the databases are continuously filled with missing values. They can have several origins: different sets of questions have been given to the students, the students did not have time to answer the question, they chose to skip it, etc. We choose to focus on this part.

3. EXPERIMENTAL

3.1 Description of experiments

The research question can be formulated as follows: how much data, created by a given number of students answering to a given number of questions is required to obtain a good estimation of the parameters of the questions, given that the students may not answer all the questions of the database? In this study, we do not aim at measuring the goodness of fit of the model, since in the first axis we know that the model is the good one: the data were simulated thanks to it.

3.2 Methods

In that study, we chose to use simulated data. This allows us to know the latent trait of the students and the coefficients of the questions, thus we are able to compare precisely the theoretical coefficients with the experimental ones. Moreover, since the data is simulated thanks to the model which will be applied, there is no interference of model misfit.

Data has been simulated for 50, 100, 500, 1000, 2000, and 3000 students, on 4, 8, 16 and 32 items, which make a total of 24 situations. The abilities of the students follow a standard normal distribution, in this we follow the examples of (Kim, Moses, & Yoo, 2015); (Neel, 2004); (Yavuz & Hambleton, 2017). The discriminations of the items follow a uniform distribution between 0.8 and 1.8, in this we follow the examples of (Svetina et al., 2013); (Yavuz & Hambleton, 2017). The difficulties of the items follow a standard normal distribution, in adequacy with the abilities, in this follow the examples of (Haberman, Sinharay, & Chon, 2013); (Svetina et al., 2013).

For a student S_j and an item Q_i this probability P_{ij} is computed by equation 1 and compared to random number computed following a uniform law between 0 and 1. If it is above, the student answered the question correctly, otherwise it is false.

The parameters have been computed thanks to the package **mirt** in R, with an “itemtype” selected at “2PL” which refers to the 2-Parameters Logistic model.

3.3 Data cleaning

The parameters of a question cannot be evaluated if it has never been answered, or if all the students answered the same thing (i.e. if they all succeeded or they all failed to that question). The data is not simulated to avoid that situation, since it could introduce bias. Instead, we remove the question of the database: in IRT terms, this kind of question is useless because it does not add any information.

3.4 Cases studies

The study has been separated in two cases.

Case A. We have full data, which means that all the students answered all the question, there is no missing value. This case is designed to be the baseline, the “perfect case”.

Case B. The students only answered half of the questions. Each student answers a to a different random subset of questions, without checking the number of student who already answered the question, nor the difficulty or discrimination of the questions.

3.5 Indicator

The results are shown in the following figs. 1 to 4. The indicator is the RMSD between the experimental and theoretical values of the questions' coefficients, which one wants as low as possible.
In accordance with the literature, we chose the value 0.3 as the threshold for the RMSD (Şahin & Anıl, 2016). In the following plots, it is represented by a bar.
We represent the results of the difficulty and discrimination parameters for the two cases A and B.

4. RESULTS AND DISCUSSION

4.1 Experiment of axis 1

4.1.1 Results

The results of case A are shown in figs. 1 and 2. The results of case B are shown in figs. 3 and 4.

4.1.2 Discussion

In the two cases, we can see that the difficulty parameters are always easier to compute than the discrimination ones. This is a phenomenon frequently noticed in the literature. As Svetina et al. (2013) points out, the RMSD of the difficulty would have been bigger if we had chosen $b \rightsquigarrow N(0, 2)$ instead of $b \rightsquigarrow N(0, 1)$ because of the imprecisions “in the long tail”, i.e. for low or high difficulties.

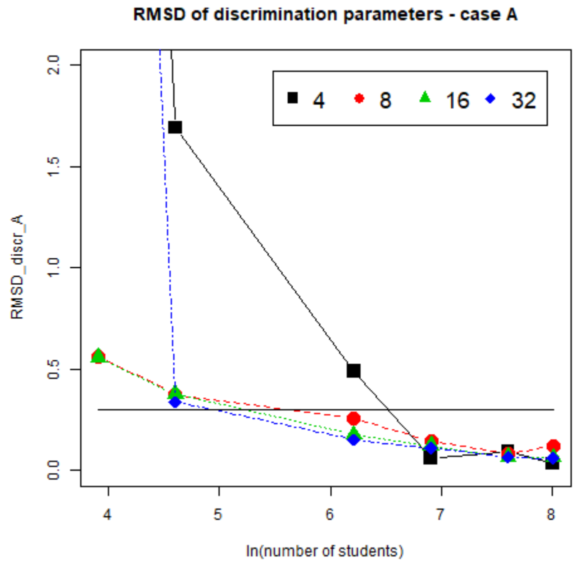

Figure 1: RMSD of discrimination parameter in case A

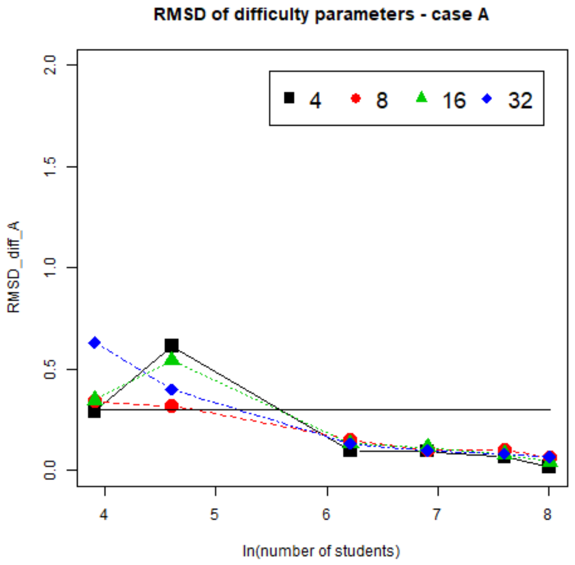

Figure 2: RMSD of difficulty parameter in case A

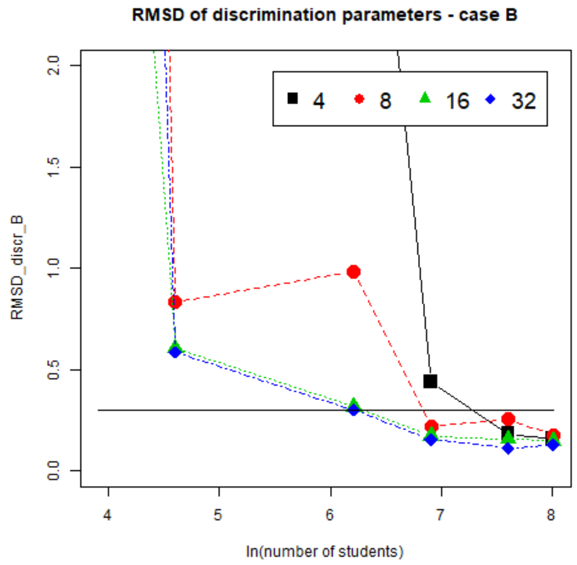

Figure 3: RMSD of discrimination parameter in case B

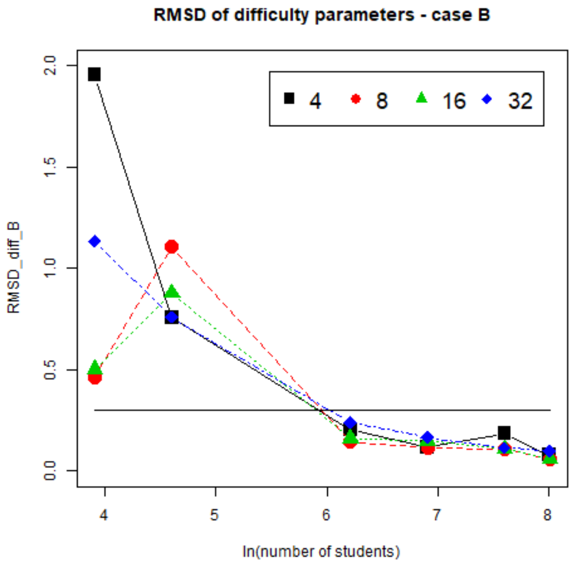

Figure 4: RMSD of difficulty parameter in case B

In case A, we confirm the findings of the literature and extend it to that package, which states that from 500 students and 8 questions the results are good. However with only 4 questions, the results are a little bit weak, as can be seen with the discrimination parameter.

Case B highlights that both the number of students and the number of questions are important parameters. This is less noticeable in case A because it converges towards acceptable situation too quickly. In case B, it can be noticed by looking at the curve representing the RMSD and the correlation of the discrimination parameters.

Case B brings out that the relationship between the percent of answers and the amount of data required might be linear. Here we only have half of the data, and for the same number of student we need twice as much questions to obtain the same quality of results, and the same holds for the situation with the same number of questions, twice as much students are required to obtain the same quality of results.

4.1.3 Conclusion

The main lesson is that when we deal with a database with missing values, we cannot use the thresholds of the literature obtained with a full database.

5. CONCLUSION AND FUTURE WORK

In this study, we aimed at understanding the effects of missing values on the reliability of parameters estimation, and the threshold of data amount. We highlighted that missing data is a parameter that has to be taken into account when one uses a database, and that the thresholds of the literature obtained with a full database cannot be used.

When facing case A, we recommend to have at least 4 questions with 1000 students, or 4 questions with 500 students; when facing case B, we recommend to have at least 8 questions with 1000 students, or 4 questions with 2000 students.

To complete this study, we will go through other cases of missing data, and use other indicators, such as correlation. That study made the hypothesis that the data respect exactly the model: we will investigate the influence of noisy data. One could also compare these results with other programs, whether other libraries in R or software such as WINSTEP or PARSCALE.

6. ACKNOWLEDGMENTS

The authors wish to thanks Charles Tijus and Simon Lemerle for their help.

This study has been partially financed by the National Association of Research and Technology of France, and Domoscio, which we thank too.

This work has been partially supported by the Spanish ministry of Economy and Competitiveness and the European Regional Development Fund, grant TIN 2017-83445-P.

References

Chuah, S. C., Drasgow, F., & Luecht, R. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education*, *19*(3), 241–255. Retrieved from http://www.tandfonline.com/doi/abs/10.1207/s15324818ame1903_5

Gao, F. & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in education*, *18*(4), 351–380. Retrieved from http://www.tandfonline.com/doi/abs/10.1207/s15324818ame1804_2

Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika*, *78*(3), 417–440. Retrieved from http://link.springer.com/article/10.1007/s11336-012-9305-1

Kim, S., Moses, T., & Yoo, H. H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*, *52*(1), 70–79. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/jedm.12063/full

Neel, J. H. (2004, November). A New Goodness-of-Fit Test for Item Response Theory. *Journal of Modern Applied Statistical Methods*, *3*(2), 581–593. doi:10.22237/jmasm/1099268760

Şahin, A. & Anıl, D. (2016, December). The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Educational Sciences: Theory & Practice*, *17*(1). doi:10.12738/estp.2017.1.0270

Svetina, D., Crawford, A. V., Levy, R., Green, S. B., Scott, L., Thompson, M., . . . Kunze, K. L. (2013). Designing small-scale tests: A simulation study of parameter recovery with the 1-PL. *Psychological Test and Assessment Modeling*, *55*(4), 335. Retrieved from https://pdfs.semanticscholar.org/c2f5/0394c4d75b98d7b7ccab91d9d429c
pdf

Wyse, A. E. & Babcock, B. (2016, June). How Does Calibration Timing and Seasonality Affect Item Parameter Estimates? *Educational and Psychological Measurement*, *76*(3), 508–527. doi:10.1177/0013164415588947

Yavuz, G. & Hambleton, R. K. (2017, April). Comparative Analyses of MIRT Models and Software (BMIRT and flexMIRT). *Educational and Psychological Measurement*, *77*(2), 263–274. doi:10.1177/0013164416661220