Item Response Theory (IRT) and large scale learning assessment in Brazil

Ruben Klein
Fundação Cesgranrio , Rio de Janeiro, Brazil
ruben@cesgranrio.org.br

This paper reviews the use of IRT models in large assessments in Brazil since 1995. It describes the "full information" equating methodology based in the use of multiple groups. The paper points out problems, new developments and challenges.

Key Words: Equating, Item Response Theory, Large scale assessment

1. Introduction

Item Response Theory (IRT) started to be used in large scale learning assessments in Brazil in 1995, in the realization of SAEB (Sistema de Avaliação da Educação Básica – Basic Education Assessment System). It is a national assessment conducted by INEP/MEC (Instituto Nacional de Estudos Educacionais Anísio Teixeira/Ministry of Education) in Reading and Mathematics at the 5th and 9th grade of Educação Fundamental, EF, (end of primary and middle school, respectively) and at the 3rd grade of Ensino Médio, EM, (High School). SAEB is realized at every two years since then.

Brazil has switched, from an EF of 8 grades to an EF of 9 grades in the last decade, introducing an earlier grade for 6 years old children. Before this change, children started EF at 7 years old. We will use, in the paper, the notation of the EF of 9 grades.

Originally the students, from public and private schools, in all grades, were sampled. But, since 2007, all students from public schools at the 5th and 9th grades of EF participate in SAEB, students from private schools and of the 3rd grade of EM are still sampled. There have been changes in the universe of students, but basically, students of very small schools are excluded. The censitary part of SAEB is being called Prova Brasil. In 2005, there was an edition of Prova Brasil, separate from SAEB. Descriptions of SAEB, until 2005, can be found in Klein & Fontanive (1995) and in Fontanive & Klein (2000).

In SAEB, until 2005, classrooms were sampled and half the students were tested in Reading and the other half in Mathematics. Starting in 2007, all students were tested in both disciplines. For instance, the total number of answered items increased from 39 to 52 in the 9th grade EF and 3rd grade EM, but the number of items per discipline decreased to 26.

SAEB uses a balanced incomplete block design (BIB) and adopted a spiral distribution of the booklets to the students. Nowadays, students are being identified and the spiraling is being done by alphabetical order. This procedure assures the randomization of the students.

SAEB uses IRT to create a unique proficiency scale for all grades and all years. To allow for this, items from the 5th grade EF are included in the 9th grade EF and items from the 9th grade EF are included in the 3rd grade EM and items from one edition are included in the next edition.

Nowadays, there a lot of state and some city assessment systems and most of them are putting their results in SAEB scale with the use of SAEB items.

The assessment systems use to have also socioeconomical and cultural questionnaires for students and also questionnaires for teachers and principals, so that other analysis can be done.

In the next section, we describe how the calibration and equating of the items is being done in SAEB and in most other Brazilian assessment systems. Finally in the last section, we present problems, new developments and challenges.

2. Calibration and Equating

In SAEB, we use IRT for multiple groups, Bock & Zimowski (1996), Zimowski et al, (1996, 2003). The calibration and equating in SAEB using IRT for multiple groups is described in Klein (2003) and in the technical reports for SAEB.

In IRT for multiple groups, the parameters of the items (estimated in a unique way, including those of the common items) and the proficiency distribution of the groups are jointly estimated.

Usually, the priors for all groups are taken to be normal distributions with different means and standard deviations. To eliminate the indeterminacy of the model, the prior for one of the groups, called reference group, is taken as N(0,1). Together with the estimation of the parameters, the means and standard deviations of the prior distributions of the other populations are also estimated. In fact, all the priors are re-estimated along the iterative estimation process, but the mean and standard deviation of the reference group are kept fixed.

The proficiency scale used in SAEB, for each discipline, started in 1997, with the joint calibration of the 1995 and 1997 assessments, with 6 groups, 3 grades of 1995 and 3 grades of 1997, with the 9th grade EF of 1997 being the reference group.

For later editions of SAEB, calibration and equating are done jointly using, what we call “full information” equating. In this procedure we introduce a new paradigm, besides the estimated parameters of the items, use also the data and analysis that gave origin to the calibration.

1st case: Common items, already calibrated, belong to one or several groups, including the reference group with distribution N(0,1).

In this case, it is enough to add to the previous analysis, the new data, fixing the estimated parameters of the previous analysis and keeping the reference group with the distribution N(0,1). The joint estimation of parameters and group distributions are done with these restrictions.

For instance, in SAEB 1999, there are items of the 5th grade of 1999 common with the 5th grade of 1997. The same is true for the other grades. Then we use 6 groups, the 3 grades of 1997 and the 3 grades of 1999. As the 9th grade of 1997 was the reference group of the 1995-1997 analysis, its distribution is already N(0,1) and it is kept as the reference group in this new analysis. All parameters of the 1997 items are kept fixed.

2nd case: Common items, already calibrated, belong to one or several groups, no old group has distribution N(0,1), but the old groups are in the required scale

In this case, we need the parameters m and s to transform linearly the mean and standard deviation of the reference group to mean 0 and standard deviation 1. These parameters have to be obtained by previous analysis. We apply the same transformation to the known item parameters

Run the program fixing known transformed parameters and using the original group as reference and supposing a N(0,1) prior.

```
b <- (b-m)/s
a <- a*s
c <- c
```

After estimation, we have to apply the inverse linear transformation to go back to the original scale. The estimation of the proficiencies is done at this scale.

More generally, if there are other calibrated items in the same scale or other common items in the new test, but related to other groups, the parameters of these items should be transformed in the same way. If there are many calibrated items of one other group, the data and all items of this group may also be used in the new calibration.

Example 1. SAEB 2001 has common items with SAEB 1999 the same way as SAEB 1999 had to SAEB 1997. But no group of SAEB 1999 has distribution N(0,1), although their items are in SAEB scale

We take the 8th grade of 1999 as reference. The mean m9908 and standard deviation sd9908 of this distribution as computed in the 1997-1999 analysis are the parameters we need to transform the mean and standard deviation of 1999 8th grade to 0 and 1, respectively.

We do the linear transformation for all item parameters of 1999

b9901 -> (bsaeb-m9908)/s9908
a9901 -> asaeb*s9908
c9901 -> csaeb

Run the program with the 3 groups of 1999 and the 3 groups of 2001, using the 9th grade of 1999 as the reference group with prior N(0,1) and keeping the 1999 transformed item parameters fixed.

Then we apply the inverse linear transformation to return the item parameters to SAEB scale and estimate the proficiencies in this scale.

bsaeb = s9908*b9901 + m9908
asaeb = a9901/s9908
csaeb = c9901.

Example 2. Equate SAEB 2003 from common items in SAEB 2001.

We need the parameters to transform linearly the mean and standard deviation of the reference group, 8th grade of 2001, SAEB scale, to 0 and 1, respectively. They can be obtained from the composition of the linear transformation to transform SAEB 2001, SAEB scale, to the scale 1999-2001(from example 1) and the linear transformation to make the distribution of the group 8th grade of 2001 in the scale 1999-2001 to have mean 0 and standard deviation 1, using the computed means of the distribution in the calibration-equating 1999-2001.

In this way, in all situations, knowing the previous analysis, we can always obtain the necessary transformation parameters.

For instance, in a State assessment of a 5th grade, with the use of common items of some SAEB 5th grade. Use as the group of reference, SAEB 5th grade.

Transform all items of SAEB 4th grade so that mean and standard deviation of 4th grade is 0 and 1, respectively. Calibrate jointly all items of the two groups. Make inverse transformation. Estimate the proficiencies.

The program we have been using is Bilog-MG, Zimowski et al (1996, 2003). Initially, we have used the DOS version. In this version, we “fix” the items using tight priors for the item parameters. In the Windows version, the parameter “c” is really fixed, when we use the PRNAME keyword in the GLOBAL command to give the file with the fixed parameters and the FIX keyword in the TEST command to give the fixed items. The program still uses tight parameter priors for the other parameters. We have used the default prior distributions N(0,1) for all groups. The mean and standard deviation of the groups are given in the output .PH2. The ideal situation is the one in which all parameters are really fixed.

Initially, we let the parameters drift, that is, we used the final parameter estimates, even for the common items. Nowadays, we keep the parameters of the common items really fixed, that is, after the final calibration-equating, we use the original fixed item parameters.
In all calibration-equating processes, a study of Differential Item Functioning (DIF) is done for common items among the groups. Initially, for items with DIF, but with separate good fitting, we considered the item as two different items, one for each group. Nowadays, we abandon the item in one of the groups. Usually, we abandon it in the upper grade, since it comes from the lower grade. In the same way, we abandon it in the new test, since it comes from the old test.

It is important to realize that the mean of the estimated proficiencies, by EAP method, of the individuals of a group will be equal to the estimated mean of that group if and only if the prior distribution used is the empirical distribution estimated at the end of the estimation. This is so since at each step of the estimation iterative procedure, the prior distribution is re-estimated and the estimated mean of the group is the mean of the posterior means (EAP) of the individuals, relative to the estimated prior. It will not be equal if some other estimation method for the proficiencies is used.

The estimation of proficiencies is done after the item calibration process. We use the EAP method with the default of the Bilog-MG, that is, we use the prior N(0,1) for all cases in all groups. This has the advantage that we know which prior to use in any new case and that if individuals in different groups have the same answers in the same test they have the same estimated proficiency. But it also has disadvantages; the mean of the estimated proficiencies for each group is not the same as the computed mean in the .PH2 output and the estimated proficiency suffers from the shrinking effect of the EAP estimator towards the mean of the prior, especially for the non-reference groups. In these senses, the empirical priors obtained in the program would be better, but they suffer from the comparison problem, which prior to use for new individuals, not belonging to any of the considered groups, and if we had the same answers in the same test for individuals in different groups, we would have different estimated proficiencies.

This methodology offers new opportunities. We can reweight the weights in each group so that we have the approximately the “same number” of answers in each common item or we may let one group have more answers in an item than another, as for instance, in a situation of real testing and pre-testing used for linking purposes.

3. Conclusion

We have described a new method of equating, using IRT for multiple groups, used in Brazil since 1999 and have pointed out the problem of not having a program that really fixes all item parameters. Another issue not discussed in the paper is how to choose common items for equating, a recommendation found in Zimowski et al (1996, p. 20) is that “common items should have relatively high discrimination power, middle range difficulty, and should be free of any appreciable DIF effect”.
Another issue is how to have comparable tests, since estimated proficiencies depend on the test, see Klein (2013).
New challenges are appearing in Brazil, since there is a trend to identify students and compute their proficiencies with precision. This involves adequate tests to individual students that possibly can best be tackled by computerized adaptive testing. We need studies to show if it is possible to use the same scale for these new tests or if it will be necessary to create a new one. Another issue is how to equate new items in this new setting.

References

Bock, R. D.& Zimowski, M.F. Multiple Group IRT (1996). In: Linden, W. J. & Hambleton, R.K. (Ed.). *Handbook of modern item response theory*. New York: Springer.

Fontanive, N.S. & Klein, R. (2000). Uma Visão sobre O Sistema de Avaliação da Educação Básica do Brasil – SAEB. Ensaio: Avaliação e Políticas Públicas em Educação, v.8, n.29, p.409-442.

Klein, R, & Fontanive, N.S. (1995). Avaliação em Larga Escala: uma proposta inovadora. Em Aberto, INEP/MEC, Vol. 15, nº 66, pp. 29-34.

Klein, R. (2003). Utilização da Teoria de Resposta ao Item no Sistema Nacional de Avaliação da Educação Básica (SAEB). Ensaio: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v.11, n.40, p. 283-296.

Klein, R. (2013). Alguns aspectos da Teoria de Resposta ao Item relativos à estimação das proficiências. Ensaio: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v.21, n.78, to appear.

Zimowski, M.F.; Muraki, E.; Mislevy, R.J.; & Bock,R.D. (1996). Bilog-MG. Scientific Software International.

Zimowski, M.F.; Muraki, E.; Mislevy, R.J.; & Bock,R.D. (2003). Bilog-MG for Windows. Scientific Software International.