TEST EQUATING WITH EVOLVING LATENT ABILITY

Silvia Bacci[1], Bruno Bertaccini[1], Carla Galluccio[1],
Leonardo Grilli[1] and Carla Rampichini[1]

[1] Department of Statistics, Computer Science, Applications "G. Parenti", (e-mail: silvia.bacci@unifi.it, bruno.bertaccini@unifi.it, carla.galluccio@unifi.it, leonardo.grilli@unifi.it, carla.rampichini@unifi.it)

ABSTRACT: In large-scale assessments, students' ability is usually evaluated using multiple test forms, which require the use of several items. In this context, calibrating items before the official tests can be difficult for different reasons. A solution is to calibrate items during the first test administration and then use these estimates in the subsequent ones. However, this approach does not consider that the populations could be significantly different in terms of average ability, which is particularly problematic when the final output of this process is a merit ranking. Our findings show that, on one side, calibrating item parameters on populations with differences in ability does not affect the final merit ranking and, on the other side, the differences in item parameter estimates are significant.

KEYWORDS: large-scale assessment, bifactor model, test equalisation

1 Introduction

In large-scale assessments, it is common practice to construct multiple test forms so as to increase test security and allow for tests to be implemented on different exam dates and times (van der Linden & Adema, 1998). This organisation requires using many items (in order to construct parallel versions of the same test for each group) and the application of equalisation methods to make the scores obtained on different test forms comparable, a relevant concern when the final output of this process is a merit ranking. Additionally, field trials are usually required to calibrate test items when using IRT models (Hambleton *et al.*, 1991) to assess subjects' ability in large-scale assessments.

It is worth understanding that calibrating items before official tests could be problematic for different reasons. Among these, the set of items used is usually not large enough to give the possibility of using in advance items that should then constitute the official tests. A possible approach is to calibrate items on the first group of subjects and then use the item parameter estimates to assess the ability of students who are subsequently administered the tests.

A problem that might arise in some contexts, such as university entrance tests, is that the population on which items are calibrated at the baseline may differ significantly from those at subsequent administrations. For example, it is reasonable to assume that students who took the test at the baseline have lower abilities than those who took it later, at least because they had more time to study and became familiar with the type of test.

The present work aims at answering the following research questions:

RQ1 Is tests equalisation, and consequently the merit ranking, affected by differences in the population average ability?

RQ2 How does calibrating items on a certain population affect estimates of ability in a population that differ for the average ability levels?

2 Statistical Model

In certain contexts, the structure of the test is characterised by the presence of subsets of questions concerning the same topic (referred to as testlets), which implies a violation of the hypothesis of local independence of the items (van der Linden & Hambleton, 2013). Thus, models capable of managing the multidimensionality of the latent trait are required.

In multidimensional item response theory (Reckase, 2009), the bifactor (BF) model (Holzinger & Swineford, 1937) is often used due to its good performance on different kinds of data. In the BF model, a common (i.e., generic, primary) latent variable is assumed to underlie all test items. In addition, specific latent variables (one for each testlet) account for the residual dependence remaining after considering the primary latent construct and due to the presence of the testlets. Primary and specific latent variables are orthogonal.

Let us consider a set of individuals $i = 1,\dots,n$ taking a test with $j = 1,\dots,J$ items divided into $s = 1,\dots,S$ sections. In the two-parameter BF model for dichotomous items Y_{ijs}, the probability that test taker i correctly answer item j of section s is defined as

$$P(Y_{ijs} = 1|\theta_{0i}, \theta_{si}) = \frac{1}{1+\exp(-[d_j + a_{0j}\theta_{0i} + a_{sj}\theta_{si}])},$$

where θ_0 is the primary latent variable, θ_s is the s-th specific latent variable, d_j denotes the difficulty parameter of item j, a_{0j} and a_{sj} represent the discrimination parameters of item j on the primary and specific constructs, respectively. If item j loads on specific factor s, $a_{sj} \neq 0$, otherwise $a_{sj} = 0$.

3 Simulation Study

To answer the two research questions RQ1 and RQ2, we performed a simulation study. A test with 50 dichotomously-scored items was generated for the study, including four testlets composed of 7, 15, 15, and 13 items, respectively. Parameters a_{0j} were sampled from a log-normal distribution $logN(0, 0.5)$ constrained to $[0.5, 2]$. Moreover, for each testlet parameters a_{sj} were sampled from a uniform distribution $[0.5, 0.7]$, corresponding to a moderate degree of local dependence between items. Difficulty parameters d_j were sampled from a normal distribution $N(0, 1)$. We assume that the same set of items is administered in two different time occasions.

The generic and the specific latent abilities θ_0 and θ_s were generated from a mixture of two independent Gaussian distributions:

$$f(\theta) = \pi_A f_A(\theta) + \pi_B f_B(\theta)$$

where $f(.)$ is the normal density and π_A and π_B are the mixture component weights, with $\pi_A + \pi_B = 1$. The mean of the mixture is $\mu_M = \pi_A \mu_A + \pi_B \mu_B$, and its variance is $\sigma_M^2 = \pi_A \sigma_A^2 + \pi_B \sigma_B^2 + \left[\pi_A \mu_A^2 + \pi_B \mu_B^2 - (\pi_A \mu_A + \pi_B \mu_B)^2\right]$.

We assume the mixture components f_A and f_B have mean $\mu_A = -2$ and $\mu_B = 2$ respectively, and common variance $\sigma^2 = 1$. We simulate two groups of subjects with different ability distributions: the baseline group (group 1) with 80% of subjects from the first component and 20% from the second one, and a second time occasion group (group 2) with 20% of subjects from the first component and 80% from the second one. Note that with this configuration, the mixture distributions of groups 1 and 2 have different means but equal variance. For each group, $N = 10,000$ response patterns were simulated. In addition, a set of 500 subjects was assumed to repeat the test, and thus, they are present in both groups, with an ability improvement of 0.5 in group 2 compared to group 1. Parameters estimation was carried out through the EM algorithm implemented in the R package `mirt`.

4 Results

To investigate RQ1, we considered the merit ranking obtained by estimating a BF model under three different strategies: (i) considering the two groups separately; (ii) considering the two groups together; (iii) using for the second group the item parameters estimated on the first one. The merit ranking resulting from each strategy was compared to the true ranking by using the Pearson correlation coefficient.

The correlation coefficients are equal to 0.86, 0.96 and 0.96, respectively, showing no differences in estimating subjects' abilities θ_0 for the two groups together or using the parameters estimated on the first groups in the second one in terms of merit ranking. Conversely, the coefficient obtained when the two models are estimated separately (strategy *i*) is remarkably lower. This result is in line with the literature on equalisation methods with non-equivalent groups.

To answer RQ2, we compared some constrained BF models. We first assessed a base (unconstrained) model (Model 0), in which the 30% of items in each testlet were in common and the other ones were considered as different, so that different parameters were estimated for the same item administered in the two time occasions. Then, four models (Model 1-4) nested in the base one were estimated, where the items within each testlet were constrained to have equal parameters across the two time occasions. We compared the constrained models with the base one using BIC, AIC, and the log-likelihood. Results provide evidence in favor of the base model, recognising an effect on item parameter estimation when populations present remarkable differences in ability.

5 Conclusions

Preliminary results above presented advice against separately calibrating tests administered in different occasions and outline the presence of an effect of populations with different ability distributions on the item parameters. Future work will focus on extending the simulation study to more general scenarios, such as different mixtures of populations and tests with only a sub-set of common items.

References

HAMBLETON, R.K., SWAMINATHAN, H., & ROGERS, H.J. 1991. *Fundamentals of Item Response Theory*. Vol. 2. Sage, California.

HOLZINGER, K.J., & SWINEFORD, F. 1937. The bi-factor method. *Psychometrika*, **2**, 41–54.

RECKASE, M.D. 2009. *Multidimensional Item Response Theory Models*. Springer, New York.

VAN DER LINDEN, W.J., & ADEMA, J.J. 1998. Simultaneous assembly of multiple test forms. *Journal of Educational Measurement*, **35**(3), 185–198.

VAN DER LINDEN, W.J., & HAMBLETON, R.K. 2013. *Handbook of Modern Item Response Theory*. Springer Science & Business Media, New York.