UNIVERSITÀ DEGLI STUDI DI UDINE

DIPARTIMENTO DI SCIENZE ECONOMICHE E STATISTICHE

A Test for the Detection of Scale Drift

Michela Battauz

December 2017

n. 7/2017

Statistica Metodologica e Sperimentale

A Test for the Detection of Scale Drift

Abstract

This paper proposes a statistical test for the detection of scale drift base on item response theory methods. When the item parameters are estimated separately for different forms of a test, they are expressed on different measurement scales. It is possible to convert them to a common metric using two constants, called equating coefficients. The equating coefficients can be estimated for two forms with common items, or derived through a chain of forms. When two forms can be linked through more than one path, each of them yields a different scale conversion. The proposal of this paper is a statistical test of whether the scale conversions deriving from different paths are equal. The approach is illustrated through a simulation study.

Index terms: equating, item response theory, linking, scale drift, scale stability, Wald test.

Introduction

Many testing programs involve several administrations over time, and the comparability of the scores is certainly an essential requirement. To this end, the equating procedures (Kolen & Brennan, 2014) can be used to adjusts for differences in difficulty across different test forms. However, various sources of variability can lead to scale drift (Haberman & Dorans, 2009), and the scores on the forms can no longer be used interchangeably. The sources of variability include both systematic and random error, and only systematic error can induce scale drift (Haberman & Dorans, 2009). Several contributions in the literature testify the importance attributed to the detection of scale drift. Many works are based on the comparison of the equated scores. Petersen et al. (1983) investigated the presence of scale drift comparing the scores of the base form to the equated scores through a chain of equatings. Puhan (2009) compared the equated scores deriving from two parallel chains, and used the notion of *difference that matters* to define a threshold beyond which to consider the differences not negligible. Liu et al.

(2009) compared the original raw-to-scale conversion to a new conversion obtained by readministration of an old form. Li et al. (2012) compared the equated scores obtained through the same chain of forms with two different equating procedures, called direct and indirect equating in the paper. Other works considered the mean scale score. Haberman et al. (2009) analyzed the effect of the year and the mouth on mean scale scores. Lee & von Davier (2013) applied quality control techniques for time series data to mean scores. Lee & Haberman (2013) proposed a regression analysis of mean test scores.

While approaches based on the equated scores have the drawback of taking in consideration many values, analyzing only the mean score involves a loss of information. This paper proposes a novel approach, which builds on the work of Battauz (2013). In particular, that paper introduced the chain equating coefficients, which are two constants that can be used to convert the item parameters from the scale of one form to the scale of another form linked through a chain of forms. The chain equating coefficients are computed as a function of the direct equating coefficients between two forms with common items. A similar derivation can be found also in (Li et al., 2012). When two forms can be linked through more than one path, each of them yields a different scale conversion. The differences can be due to random variability or systematic error. The proposal of this paper is a statistical test of whether the scale conversions deriving from different paths are equal. The procedure has the advantage of producing a single test statistic, without any loss of information. The first step to compute the equated scores, both using the true score equating or the observed score equating methods (Kolen & Brennan, 2014), is the conversion of the item parameters to a common metric using the equating coefficients. Hence, if there are no differences in the scale conversions deriving from different paths, the equated score do not present differences as well. Another advantage of the procedure is that it detects only systematic error, taking into account the presence of random error in the data. In the next section, the procedure will be described in detail. The performance of the test will then be assessed through simulation studies. The last section contains some concluding remarks.

A Test for Scale Drift

In a 3-parameter logistic (3PL) model, the probability of a correct response to item j for a subject with ability θ is given by

$$p_j(\theta; a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp\{Da_j(\theta - b_j)\}}{1 + \exp\{Da_j(\theta - b_j)\}}, \tag{1}$$

where a_j, b_j and c_j are the discrimination, difficulty and guessing parameters. The 2-parameter logistic (2PL) model is obtained when the guessing parameters c_j are set to zero, while the 1-parameter logistic (1PL) model requires also that the discrimination parameters are equal to 1. The item parameters are generally estimated by means of the marginal maximum likelihood method (Bock & Aitkin, 1981). Due to identifiability issues, the ability values are assumed to have zero mean and variance equal to one. For this reason, when the item parameters are estimated separately for different groups of individuals, the item parameter estimates are expressed on different measurement scales (Kolen & Brennan, 2014). In order to obtain comparable values, it is necessary to convert the item parameter estimates to a common metric by means of an IRT equating method (Kolen & Brennan, 2014). More specifically, it is necessary to estimate the equating coefficients, which are two constants used to perform the transformation of the item parameters. Let $A_{g-1,g}$ and $B_{g-1,g}$ be the equating coefficients between Forms $g-1$ and g. The conversion of the item parameters from the scale of Form $g-1$ to the scale of Form g is given by the following equations:

$$a_g = \frac{a_{g-1}}{A_{g-1,g}}, \qquad b_g = A_{g-1,g}\, b_{g-1} + B_{g-1,g}.$$

The methods proposed in the literature to estimate the equating coefficients, as the mean-sigma (Marco, 1977), the mean-mean (Loyd & Hoover, 1980), the mean-geometric mean (Mislevy & Bock, 1990), the Haebara (Haebara, 1980) and the Stocking-Lord (Stocking & Lord, 1983) methods, require some items in common between the forms to be linked. When two test forms can be linked through a chain of form, it is possible to compute the chain equating coefficients (Battauz, 2013). Let $p = \{1, \ldots, l\}$ be the path from Form 1 to Form l. The chain equating coefficients are given by

$$A_p = \prod_{g=2}^{l} A_{g-1,g}, \qquad B_p = \sum_{g=2}^{l} B_{g-1,g}\, A_{g,\ldots,l}\,,$$

where $A_{g,\ldots,l} = \prod_{h=g+1}^{l} A_{h-1,\,h}$ is the coefficient that links Form g to Form l, while $A_{g-1,g}$ and $B_{g-1,g}$ are the equating coefficients between Forms $g-1$ and g.

When two forms are linked through more than one path, it is possible to compare the different scale conversions deriving from each path to investigate the presence of scale drift. If the IRT model holds perfectly and the true item parameters are constant over different administrations, the equating coefficients deriving from different paths differ only because of sample variability.

Thus, any difference which can not be attributed to this source of error indicates a violation of the assumptions of the model. Suppose there are P paths that link two forms, and let $A_1, \ldots, A_P$ and $B_1, \ldots, B_P$ the equating coefficients related to these paths. These paths can possibly include a direct link if the two forms present same common items. The proposal of this paper is a statistical test with null hypothesis the lack of scale drift

$$H_0 : \begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = \cdots = \begin{pmatrix} A_p \\ B_p \end{pmatrix} = \cdots = \begin{pmatrix} A_P \\ B_P \end{pmatrix}$$

against the alternative hypothesis that at least one equality in H_0 does not hold. Let $\boldsymbol{\beta} = (A_1, \ldots, A_P, B_1, \ldots, B_P)^\top$ be the vector containing all the equating coefficients, and $\hat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$. The test statistic is given by

$$W = (\mathbf{C}\hat{\boldsymbol{\beta}})^\top (\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top)^{-1}\mathbf{C}\hat{\boldsymbol{\beta}},$$

where $\mathbf{C}$ is a block diagonal matrix composed of two blocks with dimension $(P-1) \times P$ both equal to a matrix given by $(\mathbf{1}_{P-1}, -1 \cdot \mathbf{I}_{(P-1)\times(P-1)})$, $\mathbf{1}_{P-1}$ denotes a vector of ones with dimension $P-1$, $\mathbf{I}_{(P-1)\times(P-1)}$ denotes the identity matrix with dimension $P-1$, and $\boldsymbol{\Sigma}$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. Since the equating coefficients are a function of the item parameter estimates, the delta method can be exploited to compute $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \hat{\boldsymbol{\alpha}}^\top} \mathrm{acov}(\hat{\boldsymbol{\alpha}}) \frac{\partial \hat{\boldsymbol{\beta}}^\top}{\partial \hat{\boldsymbol{\alpha}}},$$

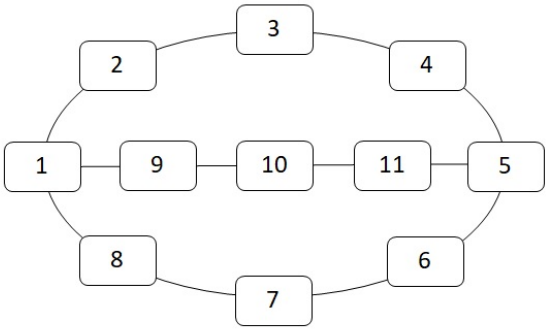
where $\hat{\boldsymbol{\alpha}}$ is the vector containing the item parameter estimates of all the forms, and $\mathrm{acov}(\hat{\boldsymbol{\alpha}})$ is the corresponding asymptotic covariance matrix. The derivatives are given in Battauz (2013). The test proposed here is a Wald test, and the asymptotic distribution of the test statistic under the null hypothesis is a Chi-square distribution with $2 \times (P-1)$ degrees of freedom. In the following section, the performance of the test will be investigated through simulation studies.

Simulation Studies

In order to investigate the performance of the method, a simulation study including various different scenarios was conducted. This study considers 11 forms, linked as shown in Figure 1. The numbers representing the forms in the figure should be regarded purely as labels and not necessarily as a sequence of time points. Forms 1 and 5 can be linked through 3 different paths, each leading to a couple of different equating coefficients. Each form

is composed of 30 items, and the number of items in common between forms directly linked is 5.

Figure 1: The linkage plan.



The ability values were generated from a normal distribution with mean and standard deviation varying across the forms. The mean was generated from a uniform distribution with range $[-0.5, 0.5]$, while the standard deviation was generated from a uniform distribution with range $[0.8, 1.2]$. The number examinees for each form is $n = \{500, 1000, 2000, 4000\}$. A 2PL model was used to generate the item responses and to estimate the item parameters. Following Battauz (2017), the difficulty parameters were generated from a standard normal distribution, while the discrimination parameters were generated from a normal distribution with mean 0.9 and standard deviation 0.3, truncated at 0.3 and 1.8. All analyses were performed using R (R Development Core Team, 2017), using the package ltm (Rizopoulos, 2006) to fit the IRT models, and the package equateIRT (Battauz, 2015a) to estimate direct and chain equating coefficients. The Haebara method was used to estimate direct equating coefficients between all forms with items in common.

In order to assess the type I error rate of the test, the first case considered does not involve scale drift. The test was applied to the equating coefficients that convert the item parameters of Form 5 to the scale of Form 1 deriving from three or only two paths. Figure 2 shows the empirical type I error rate of the test at different sample sizes. The nominal significance level was chosen to be 0.05. The empirical significance level in some cases is slightly lower than the nominal level, especially for small sample sizes. The difference can be attributed to the fact the distribution of the test statistic holds asymptotically, and also to small inaccuracies in the computation of the covariance matrix of the item parameter estimates, which is based on numerical computation of the hessian matrix.

In order to investigate the detection rate of the test (i.e. the power of the

Figure 2: Empirical type I error rate (significance level set at 0.05).

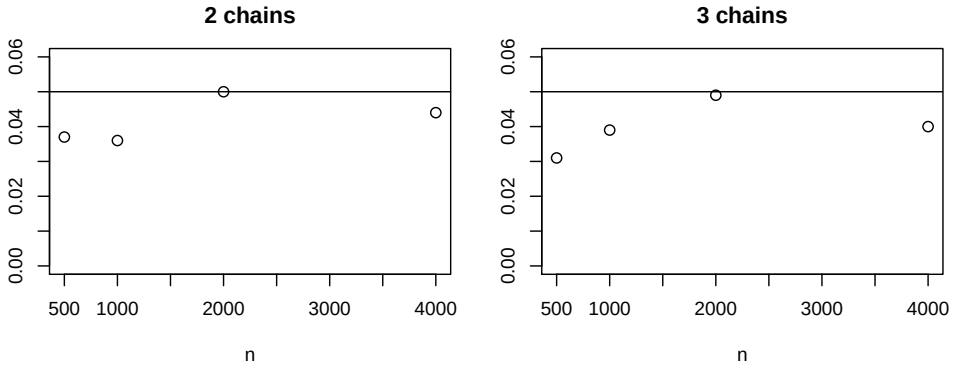


Table 1: Scale conversions from Form 5 to Form 1 obtained using the true item parameters and the Haebara method.

	equating coeffients	
	A	B
no drift	1.186	-0.756
moderate drift	1.200	-1.168
severe drift	1.518	-1.615

test), some item parameters were manipulated on purpose in order to generate a drift of the scale in one of the paths that link Forms 1 and 5. More specifically, the parameters of two items in Form 3 and two items in Form 4 were modified by adding a value of 0.4. Two cases were considered: only difficulty parameters modified, and both difficulty and discrimination parameters modified. Table 1 reports the chain equating coefficients obtained using the true item parameters and using the Haebara method for the estimation of the direct equating coefficients. Hence, these values represent the scale conversion without sample variability, so the differences are due only to scale drift. In order to have a better understanding of the magnitude of the drift, these equating coefficients were used to compute the equated scores. Since the true score equating method and the observed score equating method gave very similar results, here only the scores obtained with the latter are shown. Figure 3 represents the difference of the equated scores using the conversion with and without scale drift. When only the difficulty parameters were manipulated, the maximum difference in the equated scores was 1.9, while when both difficulty and discrimination parameters were manipulated the maximum difference was 3.2. On the basis of these values, the two cases were labeled as moderate and severe scale drift.

Figure 3: Difference between equated scores using the conversion with and without scale drift.

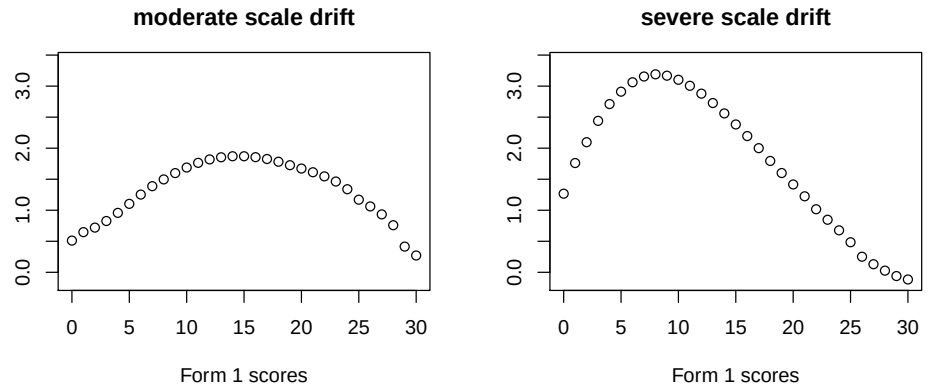
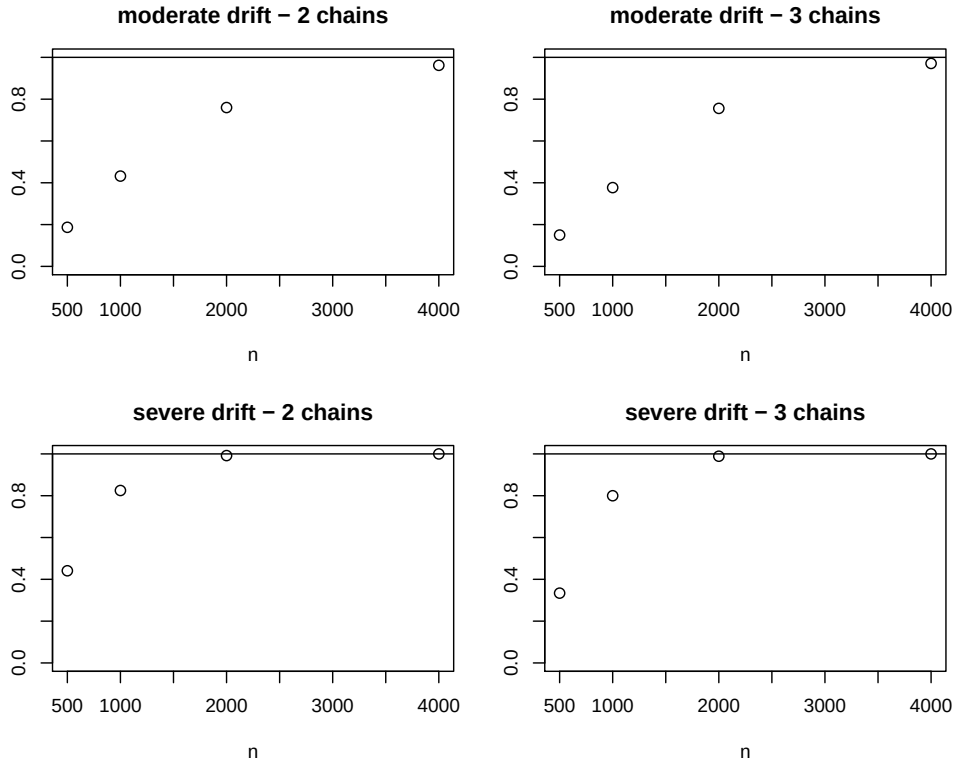


Figure 4 shows the empirical power of the test. As expected, the power increases with the sample size and it is higher when the magnitude of the scale drift is larger. The values shown in the figure can serve as an indication of the minimum sample size necessary to detect scale drift.

Conclusions

The proposals in the literature for the detection of scale drift focused on the analysis of the equated scores. Following an IRT approach for test equating, makes possible the comparison of the scale conversions. Hence, the comparison is synthesized in a single value, which is the test statistics. If the scale conversion varies across the different paths, also the equated scores will exhibit differences. However, comparing the scale conversions through a test based on the equating coefficients is more convenient, as the drift is detected at the origin. The procedure makes possible to compare more paths at one time. These paths can include a direct link and can also be partially overlapping. It is well known that the random error increases with the length of the chain (Battauz, 2015b). Hence, with longer chains, it is necessary a larger sample size to ensure a high detection rate. The standard errors of the equating coefficients, computed as explained in Battauz (2013) and implemented in the R package equateIRT, can give an indication of the amount of random variability. Treating the problem as a statistical test permits to account for the random variability of the equating coefficients and to detect only systematic differences in the scale conversion. When the sample size is very large, the random variability is very limited and the test tends to reject

Figure 4: Empirical power of the test.



the null hypothesis even if the difference in the scale conversion is very small. Thus, a comparison of the scores obtained using the different scale conversions, as shown in Figure 3, is still informative with respect to the magnitude of the drift.

References

Battauz, M. (2013). IRT test equating in complex linkage plans. *Psychometrika*, 78(3), 464–480.

Battauz, M. (2015a). equateIRT: An R package for IRT test equating. *Journal of Statistical Software*, 68(7), 1–22.

Battauz, M. (2015b). Factors affecting the variability of IRT equating coefficients. *Statistica Neerlandica*, 69(2), 85–101.

Battauz, M. (2017). Multiple equating of separate IRT calibrations. *Psychometrika*, 82(3), 610–636.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4), 443–459.

Haberman, S. & Dorans, N. J. (2009). *Scale Consistency, Drift, Stability: Definitions, Distinctions and Principles*. Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education. San Diego, CA.

Haberman, S., Guo, H., Liu, J., & Dorans, N. J. (2009). *Consistency of SAT I: Reasoning Test Score Conversions*. Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education. San Diego, CA.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22, 144–149.

Kolen, M. & Brennan, R. (2014). *Test Equating, Scaling, and Linking: Methods and Practices (3rd ed.)*. New York: Springer.

Lee, Y.-H. & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika*, 78(4), 815–829.

Lee, Y.-H. & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika*, 78(3), 557–575.

Li, D., Jiang, Y., & von Davier, A. A. (2012). The accuracy and consistency of a series of IRT true score equatings. *Journal of Educational Measurement*, 49(2), 167–189.

Liu, J., Curley, E., & Low, A. (2009). A scale drift study. *ETS Research Report Series*, 2009(2), i–77.

Loyd, B. H. & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, 17, 179–193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139–160.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG 3. Item Analysis and Test Scoring with Binary Logistic Models.* Mooresville, IN: Scientific Software.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics*, 8(2), 137–156.

Puhan, G. (2009). Detecting and correcting scale drift in test equating: An illustration from a large scale testing program. *Applied Measurement in Education*, 22(1), 79–103.

R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of statistical software*, 17(5), 1–25.

Stocking, M. & Lord, M. L. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.