Testing for the Presence of Scale Drift: An Example

Verifica della Presenza di Scale Drift: un Esempio

Michela Battauz

Abstract The comparability of the scores is a fundamental requirement in testing programs that involve several administrations over time. Differences in test difficulty can be adjusted by employing equating procedures. However, various sources of systematic error can lead to scale drift. Recently, a statistical test for the detection of scale drift under the item response theory framework was proposed. The test is based on the comparison of the equating coefficients that convert the item parameters to the scale of the base form. After briefly explaining the methodology, this paper presents an application to TIMSS achievement data.

Abstract *La comparabilità dei punteggi è un requisito fondamentale nei programmi di valutazione attraverso test somministrati ripetutamente nel tempo. Le differenze nella difficoltà dei test si possono correggere impiegando procedure di equating. Tuttavia, diverse fonti di errore sistematico possono portare a scale drift. Recentemente, è stato proposto un test statistico per rilevare lo scale drift nel contesto della item response theory. Il test si basa sulla comparazione dei coefficienti di equating che convertono i parametri degli item nella scala di riferimento. Dopo una breve spiegazione della metodologia, questo articolo presenta un'applicazione ai dati TIMSS sull'apprendimento.*

Key words: equating, item response theory, scale stability

1 Introduction

Students' achievement level can be monitored on the basis of large scale testing programs. The fundamental requirement to guarantee a fair evaluation is the comparability of the achievement levels over different administrations. Certainly, the

Michela Battauz
University of Udine - Department of Economics and Statistics, via Tomandini 30/A 33100 Udine (Italy), e-mail: michela.battauz@uniud.it

row scores (as for example the number of correct responses) are not directly comparable because they depend on the difficulty of the test form, which can be different across the administrations. Equating is a statistical process that adjusts for differences in difficulty of the forms of a test. The literature proposes various equating methods [8], and this paper focuses on the Item Response Theory (IRT) approach. However, various sources of variability can lead to scale drift [7], causing the scores to be not comparable. A statistical test for the detection of scale drift is proposed in [3]. In this paper, the methodology is briefly explained and illustrated through an application to TIMSS achievement data.

2 Models and Methods

The 2-Parameter Logistic (2PL) model is an IRT model for dichotomous responses. The probability of a correct response to item j is modeled as a function of the ability level, θ, and the item parameters a_j and b_j

$$P(a_j, b_j|\theta) = \frac{\exp\{a_j(\theta - b_j)\}}{1 + \exp\{a_j(\theta - b_j)\}}. \tag{1}$$

The 1-Parameter Logistic (1PL) model is special case that results when the discrimination parameters a_j are equal to one (for a broad review of IRT models see [14]). These models are typically estimated using the marginal maximum likelihood method [4], which assumes a standard normal distribution for θ. For this reason, when the parameters of the model are estimated separately for different groups of subjects, the item parameter estimates are expressed on different measurement scales. The item parameters can be converted from the scale of Form $g-1$ to the scale of Form g using the following equations

$$a_{jg} = \frac{a_{j,g-1}}{A_{g-1,g}}, \qquad b_{jg} = A_{g-1,g}\, b_{j,g-1} + B_{g-1,g}, \tag{2}$$

where $A_{g-1,g}$ and $B_{g-1,g}$ are two unknown constants called equating coefficients. The literature proposes various methods for the estimation of the equating coefficients between two forms with some common items [8]. When two forms can be linked through a chain of forms, it is possible to compute the chain equating coefficients [1]

$$A_p = \prod_{g=2}^{l} A_{g-1,g}, \qquad B_p = \sum_{g=2}^{l} B_{g-1,g}\, A_{g,\ldots,l}\,, \tag{3}$$

where $p = \{1, \ldots, l\}$ is the path from Form 1 to Form l, and $A_{g,\ldots,l} = \prod_{h=g+1}^{l} A_{h-1,h}$ is the coefficient that links Form g to Form l. Each path that links two forms yields a different scale conversion. The differences are due to sampling variability or to systematic error. Since the latter can lead to scale drift, the detection of differences in the scale conversion that can not be attributed to random error, indicates the presence of scale drift. The proposal for the detection of scale drift in [3] is a test with null hypothesis

$$H_0 : \begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = \cdots = \begin{pmatrix} A_p \\ B_p \end{pmatrix} = \cdots = \begin{pmatrix} A_P \\ B_P \end{pmatrix} \tag{4}$$

and as test statistics

$$W = (\mathbf{C}\hat{\beta})^\top (\mathbf{C}\Sigma\mathbf{C}^\top)^{-1}\mathbf{C}\hat{\beta}, \tag{5}$$

where $\hat{\beta} = (\hat{A}_1, \dots, \hat{A}_P, \hat{B}_1, \dots, \hat{B}_P)^\top$, Σ is the covariance matrix of $\hat{\beta}$, $\mathbf{C}$ is a block diagonal matrix composed of two blocks both equal to $(\mathbf{1}_{P-1}, -\mathbf{I}_{P-1})$, $\mathbf{1}_{P-1}$ denotes a vector of ones with dimension $P-1$, and $\mathbf{I}_{P-1}$ denotes the identity matrix with dimension $P-1$. The covariance matrix can be computed using the delta method, considering that the equating coefficients are a function of the item parameter estimates in different administrations. Under the null hypothesis, the test statistic follows asymptotically a Chi-square distribution with $2 \times (P-1)$ degrees of freedom.

3 An Example

To illustrate the application of the procedure we used data collected for TIMSS 2011, considering achievement data in Mathematics of students at the fourth grade in Italy. Students were administered one of 14 forms (booklets). These forms present items in common as shown in Figure 1. Only dichotomous items were considered for this analysis. The total number of examinees is 3992, distributed quite uniformly between the different forms. The number of items for each form ranges between 20 and 27, while the number of common items ranges between 8 and 14.

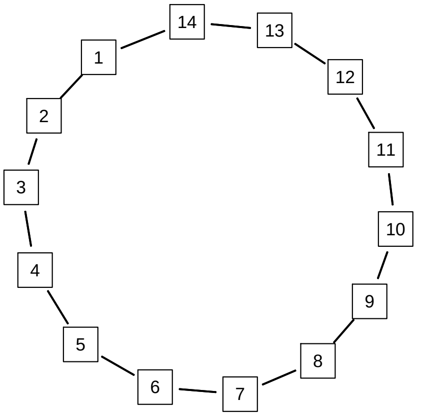

Fig. 1 Linkage plan of the example.

The 2PL model was fit to the data of each form separately. All analyses were performed using the R statistical software [13]. The mirt package [5] was used for the estimation of the IRT models, while the equateIRT [2] package was used for the estimation of the equating coefficients.

The direct equating coefficients between forms with common items were computed using the Haebara method, and the chain equating coefficients to convert the item parameters from the scale of Form 8 to the scale of Form 1 are reported in Table 1. The two paths that connect these forms present quite different equating coefficients. Anyway, the test, also reported in the table, indicates that the differences are not statistically significant at the 0.05 level.

Table 1 Estimates of chain equating coefficients (standard errors) and scale drift test.

Path	A_p	B_p
$p = \{8, 7, 6, 5, 4, 3, 2, 1\}$	1.57 (0.36)	-0.43 (0.24)
$p = \{8, 9, 10, 11, 12, 13, 14, 1\}$	0.89 (0.19)	0.02 (0.16)
$W = 4.73$, df = 2, p-value = 0.094		

4 Discussion and Conclusions

The proposal of this paper constitutes a novel approach in the literature concerned with the detection of scale drift. While traditional methods compare the scores resulting from different administrations [9, 10, 11, 12], the approach followed here is based on the comparison of the equating coefficients. This new approach permits to formulate a statistical test for the detection of scale drift, thus allowing to take into account the presence of random error.

If the test indicates that the scale conversions deriving from different paths are different, it is then necessary to investigate which items are responsible of the drift. This can be performed using tests for the detection of differential item functioning between pairs of forms [6]. After removing these items, the test for the detection of scale drift can be performed again to verify if the scale conversions can be considered equal.

References

1. Battauz, M.: IRT test equating in complex linkage plans. Psychometrika. **78**, 464–480 (2013)
2. Battauz, M.: equateIRT: An R package for IRT test equating. Journal of Statistical Software.**68**, 1–22 (2015)

3. Battauz, M.: A test for the detection of scale drift. Working paper n. 7/2017, Department of Economics and Statistics, University of Udine (2017)
4. Bock, R. D., Aitkin, M.: Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika. **46**, 443–459 (1981)
5. Chalmers, R.: mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software. **48**, 1–29 (2012)
6. Donoghue, J. R., Isham, S. P.: A comparison of procedures to detect item parameter drift. Applied Psychological Measurement. **22**, 33–51 (1998)
7. Haberman, S., Dorans, N. J.: 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 (2009)
8. Kolen, M. J., Brennan, R. L.: Test Equating, Scaling, and Linking. Springer, New York (2014)
9. Lee, Y.-H., Haberman, S. J.: Harmonic regression and scale stability. Psychometrika. **78**, 815–829 (2013)
10. Lee, Y.-H., von Davier, A. A.: Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. Psychometrika. **78**, 557–575 (2013)
11. Li, D., Jiang, Y., von Davier, A. A.: The accuracy and consistency of a series of IRT true score equatings. Journal of Educational Measurement. **49**, 167–189 (2012)
12. Puhan, G.: Detecting and correcting scale drift in test equating: An illustration from a large scale testing program. Applied Measurement in Education. **22**, 79–103 (2009)
13. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2017)
14. van der Linden, W. J.: Handbook of Item Response Theory, Volume One: Models. Chapman & Hall/CRC, Boca Raton (2016)