Received: 15 October 2019 | Revised: 31 May 2021 | Accepted: 15 July 2022

DOI: 10.1111/stan.12277

ORIGINAL ARTICLE

WILEY

Testing for differences in chain equating

Michela Battauz

Department of Economics and Statistics, University of Udine, Udine, Italy

Correspondence
Michela Battauz, Department of Economics and Statistics, University of Udine, Udine 33100, Italy
Email: michela.battauz@uniud.it

Funding information
University of Udine, Grant/Award Number: PRID 2015

The comparability of the scores obtained in different forms of a test is certainly an essential requirement. This paper proposes a statistical test for the detection of noncomparable scores based on item response theory (IRT) methods. When the IRT model is fit separately for different forms of a test, the item parameter estimates are expressed on different measurement scales. The first step to obtain comparable scores is to convert the item parameters 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. The proposal of this paper is a statistical test to verify whether the scale conversions provided by the equating coefficients are as expected when the assumptions of the model are satisfied, hence leading to comparable scores. The method is illustrated through simulation studies and a real-data example.

KEYWORDS

equating, item response theory, linking, scale drift, scale stability, Wald test

1 | 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 adjust for differences in difficulty across the test forms.

The equating methods are commonly defined for two forms, and the comparability of scores across all the forms is obtained by repeating the conversion of the scores through a chain of forms (see e.g. Kolen & Brennan, 2014, sections 4.4 and 5.5.2). Equating can be performed using either classical test theory or item response theory (IRT). Within IRT, the procedure first requires the transformation of the item parameters to a common scale and then the computation of the equated scores using the true score equating or the observed score equating methods (see Kolen & Brennan, 2014, chapter 6). The equated scores are then the equivalent number of correct answers in different test forms. The transformation of the item parameters from the scale of one form to the scale of another form involves two unknown constants, often called equating coefficients (see e.g. Ogasawara, 2001). The methods proposed in the literature for the estimation of the equating coefficients typically require that the two forms have some items in common. Following an IRT approach, Battauz (2013) derived the chain equating coefficients to convert the item parameters of two forms through a chain of forms. When two forms can be linked through more than one path, each of them yields a different scale conversion. These differences can be due to random variability or systematic error. The detection of the latter is of fundamental importance because it can induce scale drift (Haberman & Dorans, 2009) and hence the noncomparability of the scores being equated.

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, Cook, and Stocking (1983) investigated the presence of scale drift comparing the scores of the base form to the equated scores obtained through a chain of equatings. The design used is called "equating in a circle" (see Brennan & Kolen, 1987; Kolen & Brennan, 2014), which is based on the comparison of the scores on one form to the scores obtained by equating such form to itself through a chain of equatings. Puhan (2009) compared the equated scores deriving from two parallel chains, and used the notion of *difference that matters* (defined in Dorans & Feigenbaum, 1994) to define a threshold beyond which to consider the differences not negligible. Liu, Curley, and Low (2009) compared the original raw-to-scale conversion to a new conversion obtained by re-administration of an old form. Li, Jiang, and von Davier (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, Guo, Liu, and Dorans (2009) analyzed the effect of the year and the month on mean scale scores. Lee and von Davier (2013) applied quality control techniques for time series data to mean scores. Lee and Haberman (2013) proposed a regression analysis of mean test scores.

While these studies are based on the comparison of the equated scores, this paper focuses on the first step of the IRT equating process, which is the transformation of the item parameter estimates. In particular, this paper proposes a novel approach based on statistical testing of hypothesis regarding the chain equating coefficients. More specifically, the test verifies whether the equating coefficients deriving from different paths that link the same two forms are equal. Another test verifies whether the chain equating coefficients that link a form to itself differ from the identity transformation. The proposal has the advantage of detecting only systematic error, while taking into account the presence of random error in the data. Under the IRT framework, the first step to compute the equated scores, both using the true score equating or the observed score equating methods (Kolen & Brennan, 2014, sections 6.5 and 6.6), is the conversion of the item parameters to a common metric using the equating coefficients. Hence, any difference in the equating coefficients propagates to the equated scores. However, while approaches based on the equated scores require the comparison of many values, and analyzing only the mean score involves a loss of information, our proposal synthesizes the information in a single value, which is

the test statistics. It is important to highlight that the statistical tests proposed in this paper focus on the equating coefficients, which are used to convert the item parameter estimates in the first step of the equating procedure. The computation of the equated scores is a further step that is not directly involved in these tests. However, in the IRT approach, the equated scores for the same two forms could differ only because the equating coefficients differ. Hence, equal equating coefficients imply equal equated scores, while different equating coefficients could reveal the presence of scale drift.

In the next section, the procedure will be described in detail. The performance of the test will then be assessed through simulation studies and applied to a real-data example. The last section contains some concluding remarks.

2 | MODELS AND METHODS

In IRT, the probability of giving a certain response to a set of items is modeled as a function of a latent variable, here denoted by θ. In particular, the three-parameter logistic (3PL) model is used when the responses are dichotomous and it models the probability of a positive response to item j as follows:

$$p(\theta; a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp\left\{a_j(\theta - b_j)\right\}}{1 + \exp\left\{a_j(\theta - b_j)\right\}}, \tag{1}$$

where a_j, b_j, and c_j are item parameters called discrimination, difficulty, and guessing, respectively. The two-parameter logistic (2PL) model is obtained when the guessing parameters c_j are set to zero, while the one-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, section 6.2). In order to obtain comparable values, it is first necessary to convert the item parameter estimates to a common metric (Kolen & Brennan, 2014, section 6.3). To this end, it is necessary to estimate the equating coefficients, which are two constants used to perform the transformation of the item parameters. Let $A_{1,2}$ and $B_{1,2}$ be the equating coefficients between Forms 1 and 2. The conversion of the item parameters from the scale of Form 1 to the scale of Form 2 is given by the following equations:

$$a_{j,2} = \frac{a_{j,1}}{A_{1,2}}, \quad b_{j,2} = A_{1,2}\, b_{j,1} + B_{1,2},$$

while the abilities are transformed using the following equation:

$$\theta_2 = A_{1,2}\, \theta_1 + B_{1,2},$$

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 forms, it is possible to compute the chain equating coefficients (Battauz, 2013). Let $p = \{1, \dots, l\}$ be the path from Form 1 to Form l. The chain equating coefficients are given by

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

where $A_{g,\dots,l} = \prod_{h=g+1}^{l} A_{h-1,h}$ is the coefficient that links Form g to Form l. When two forms are linked through more than one path, it is possible to compare the different scale conversions deriving from each path. If the IRT model holds perfectly, the equating coefficients deriving from different paths differ only because of sample variability. Thus, any difference which cannot 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 $A_1, \dots, A_P$ and $B_1, \dots, B_P$ be the equating coefficients related to these paths. These paths can possibly include a direct link if the two forms have same common items. The proposal of this paper is a statistical test with null hypothesis the equality of the equating coefficients

$$H_0 : \begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = \cdots = \begin{pmatrix} A_P \\ B_P \end{pmatrix}, \tag{2}$$

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

$$W = (\mathbf{C}\hat{\boldsymbol{\beta}})^\top (\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top)^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}, \tag{3}$$

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 $(\mathbf{1}_{P-1}, -1 \cdot \mathbf{I}_{(P-1)})$, $\mathbf{1}_{P-1}$ denotes a vector of ones with dimension $P-1$, $\mathbf{I}_{(P-1)}$ 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 $\boldsymbol{\Sigma}$

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

where $\hat{\boldsymbol{\alpha}}$ is the vector containing the item parameter estimates of all the forms, and $\operatorname{acov}(\hat{\boldsymbol{\alpha}})$ is the corresponding asymptotic covariance matrix. The derivatives are given in Battauz (2013). The test proposed here is a Wald-type 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.

When a form can be linked to itself through a chain of other forms, the equating process is expected to return the identity transformation (Brennan & Kolen, 1987). So, an alternative test considers only one chain that links one form to itself. In this case, the transformation of the item parameters and the abilities should return the values unchanged. This means that the equating coefficients related to this path should be $A_p = 1$ and $B_p = 0$. However, the estimated equating coefficients differ from these values due to sample variability. Hence, any source of error that can

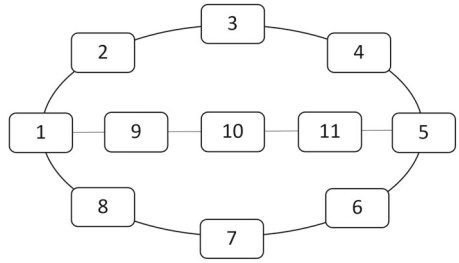

FIGURE 1 The linkage plan of the simulation study

not be attributed to sample variability reveals systematic errors in the data. The null hypothesis of the test is

$$H_0 : \begin{pmatrix} A_p \\ B_p \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{4}$$

and the test statistic is

$$C = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{5}$$

where $\hat{\boldsymbol{\beta}} = (\hat{A}_p, \hat{B}_p)^\top$, $\boldsymbol{\beta}_0 = (1, 0)^\top$, and $\boldsymbol{\Sigma}$ is the asymptotic covariance matrix of $(\hat{A}_p, \hat{B}_p)^\top$. The asymptotic distribution of the test statistic under the null hypothesis is a Chi-square distribution with 2 degrees of freedom.

In the following section, the performance of these tests will be investigated through simulation studies.

3 | 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 three 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 five.

The ability values were generated from a normal distribution with mean and *SD* that vary across the forms. The mean was generated from a uniform distribution with range $[-0.5, 0.5]$, while the *SD* was generated from a uniform distribution with range $[0.8, 1.2]$. The parameters of these distributions were in line with the values used in Ogasawara (2003).

The number examinees is constant across the forms and takes values $n = 500, 1000, 2000, 4000$. 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 *SD* 0.3, truncated at 0.3 and 1.8. All analyses were performed in R (R Development Core Team, 2021), using the package mirt (Chalmers, 2012) to fit the IRT models, and the package

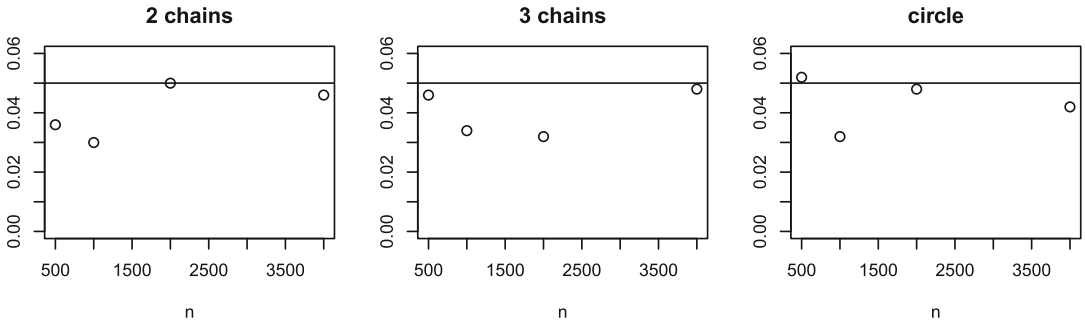

FIGURE 2 Empirical type I error rate (significance level set at 0.05)

equateIRT (Battauz, 2015a) to estimate direct and chain equating coefficients. The code developed to perform the tests proposed in this paper is publicly available in the R package equateIRT. The Haebara method was used to estimate the direct equating coefficients between all forms with items in common. The results are based on 500 simulated datasets for each setting.

In order to assess the type I error rate of the test, the first case considered does not involve differences in the equating coefficients. Test statistic (3) was computed considering three or only two paths from Form 5 to Form 1. Test statistics (5) was computed for the path that links Form 1 to itself through Forms from 2 to 8. 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 different than the nominal level. The difference can be attributed to the distribution of the test statistic, which holds only asymptotically, and also to small inaccuracies in the computation of the covariance matrix of the item parameter estimates. Figure 3 gives a more complete picture of the distribution of the test statistic. The figure shows the Q-Q plots, where the test statistics computed on the simulated datasets are compared to the theoretical quantiles of a Chi-square distribution. The empirical distribution tends to get closer to the theoretical distribution when the sample size increases, though there are some exceptions.

Similar comments apply to the case of 10 common items (all other settings kept unchanged), reported in the supplementary material that accompanies this paper. See Figure S1 for the type I error rate and Figure S2 for the Q-Q plots.

In order to investigate the detection rate of the test (i.e. the power of the test), we generated on purpose differences in the equating coefficients deriving from different paths, manipulating some item parameters in one of the paths that links 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 (labeled as "case 1"), and both difficulty and discrimination parameters modified (labeled as "case 2"). All other settings were kept unchanged. 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 that the differences are only due to systematic error. If the item parameters are not modified, the true equating coefficients are $A_{5,1} = \sigma_5/\sigma_1 = 1.186$ and $B_{5,1} = (\mu_5 - \mu_1)/\sigma_1 = -0.756$, where μ_g and σ_g denote the mean and the *SD* of the abilities in group g. They correspond to the values reported in Table 1 for the case with no drift. In order to have a better understanding of the effect of this error, these equating coefficients were used to compute the equated (summed) scores. Since the true score equating method and the observed score equating method gave very similar results, here only the scores obtained with

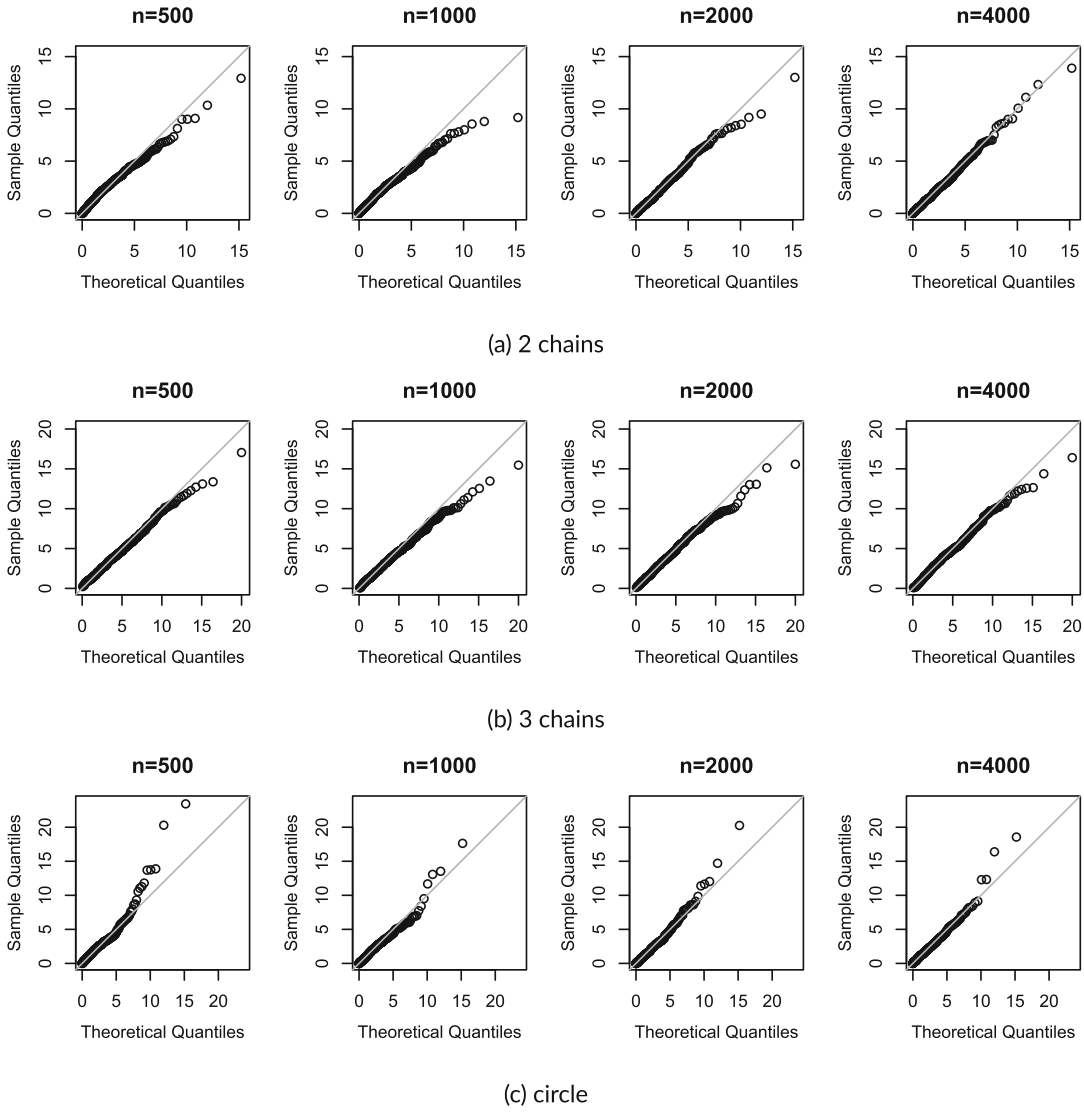

FIGURE 3 Q-Q plots of test statistics of simulated datasets. (a) Two chains; (b) three chains; (c) circle

the latter are shown. Figure 4 represents the differences of the equated scores using the equating coefficients reported in Table 1.

Figure 5 shows the empirical power of the test. Not surprisingly, the power increases with the sample size and tends to one. The values shown in the figure can serve as an indication of the minimum sample size necessary to detect differences in scale conversion in similar settings.

In order to better understand the effect of the number of common items, the number of items with drift, and the magnitude of drift, further simulation studies were run, and the results are presented in supporting information that accompanies this paper. The settings are summarized in Table S1, where Setting 1 refers to the results presented above. As expected, increasing the number of items with drift (from 2 to 4) determines a greater change in the equating coefficients (Table S2) and a more important difference in the equated scores (Figure S3). The magnitude

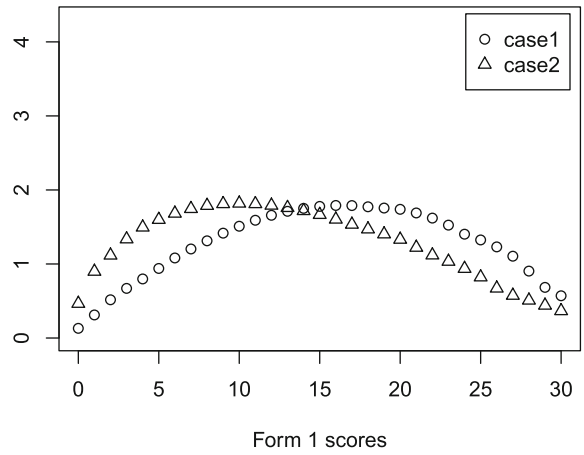

FIGURE 4 Differences between equated scores using different paths.

TABLE 1 Scale conversions from Form 5 to Form 1 using the path {5, 4, 3, 2, 1} obtained using the true item parameters and the Haebara method

	Equating coefficients	
	A	**B**
Equal equating coefficients (no drift)	1.186	−0.756
Different equating coefficients, case 1	1.121	−1.081
Different equating coefficients, case 2	1.248	−1.180

of the drift was assessed by adding a value of 0.2 to the item parameters with drift, instead of 0.4. Not surprisingly, when the drift is smaller, the difference of the equating coefficients is smaller and so is the difference of the equated scores. The effect of the number of common items is less obvious. When only the difficulty parameters are modified (case 1) increasing the number of common items from 5 to 10 leads to smaller differences of the equating coefficients and of the equated scores. However, when both the difficulty and the discrimination parameters are modified (case 2), increasing the number of common items leads to larger differences. As expected, the power of the tests is higher when the differences in the equating coefficients is larger.

4 | REAL-DATA APPLICATION

The tests proposed in this paper were applied to data collected for TIMSS 2015 (Mullis & Martin, 2013), which are openly available at https://timssandpirls.bc.edu/timss2015/international-database/. We considered the Mathematics achievement data of the 10,029 students from United States. Each student received one of 14 forms (booklets) linked as shown in Figure 6.

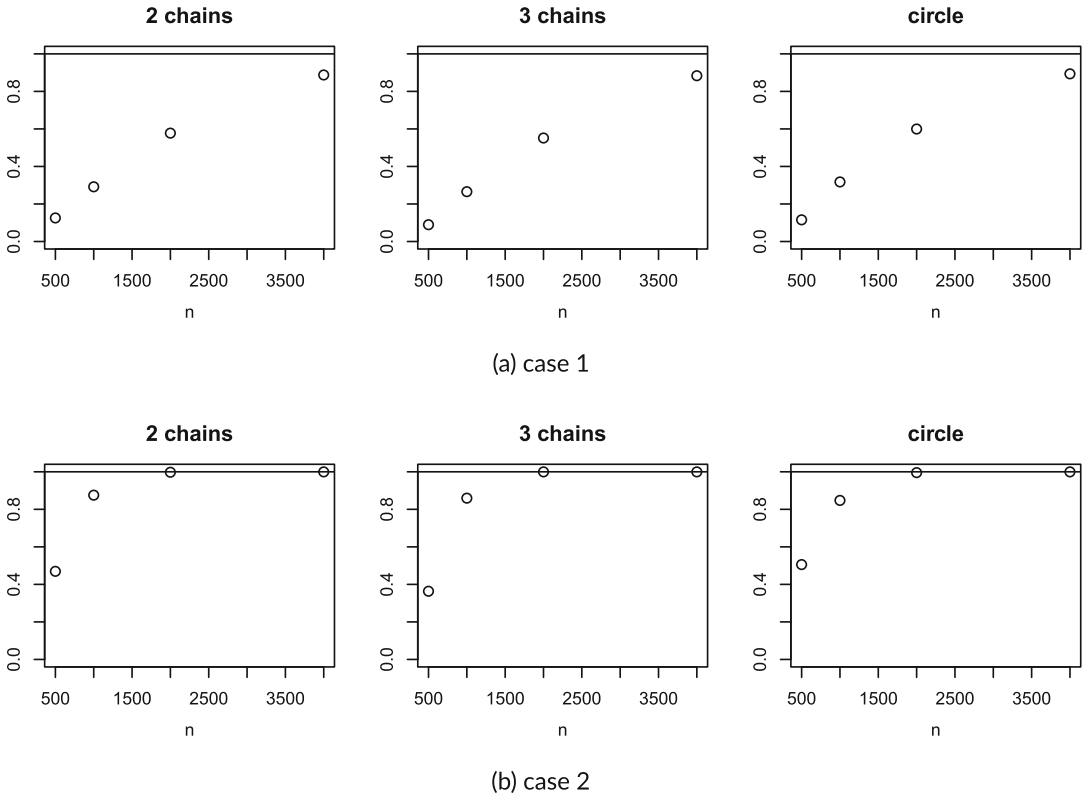

FIGURE 5 Empirical power of the test. (a) Case 1; (b) case 2

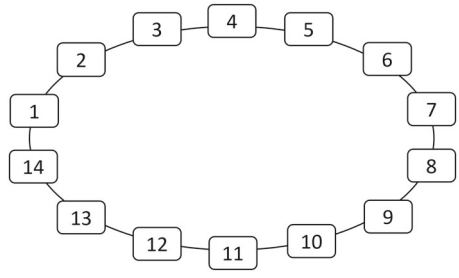

FIGURE 6 The linkage plan of the real-data application

Considering only the dichotomous items, each form is composed of a number of items ranging between 21 and 29. The number of items in common between two forms ranges between 10 and 15. The 2PL model was fit to each form separately. The chain equating coefficients that link Form 1 to Form 8 were then computed for two different paths and are reported in Table 2. These equating coefficients are quite different. Under the assumption of invariance of item parameters, the equating coefficients for path $\{1, 2, 3, 4, 5, 6, 7, 8\}$ indicate that the students that received Form 8 are on average more able than the students that received Form 1. In fact, to convert the abilities from Form 1 to Form 8 the equation is $1.08\theta_1 - 0.67$, while the abilities on Form 8 have zero mean

TABLE 2 Estimates of chain equating coefficients in the real-data application and statistical tests

path	A_p	B_p
$p = \{1, 2, 3, 4, 5, 6, 7, 8\}$	1.08 (0.14)	-0.67 (0.11)
$p = \{1, 14, 13, 12, 11, 10, 9, 8\}$	0.74 (0.09)	0.29 (0.09)
test of hypothesis (2): $W = 64.97$, df = 2, p-value <0.001		
$p = \{ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1 \}$	1.49 (0.25)	-1.37 (0.20)
test of hypothesis (4): $C = 49.14$, df = 2, p-value <0.001		

Note: *SE*s in parentheses.

and variance equal to one. However, the equating coefficients of path $\{1, 14, 13, 12, 11, 10, 9, 8\}$ indicate that the students that received Form 1 are the more able. The test indicates that the difference is statistically significant. Consistently with these findings, the equating coefficients that link Form 1 to itself through a chain composed of all the other forms are significantly different from 1 and 0 (see Table 2). A deeper look at the booklet design (Rutkowski, von Davier, & Rutkowski, 2014, p. 244), shows that the Mathematics items of each booklet are divided in two blocks and that the subsequent booklet is composed of the second block of the previous booklet plus another block. So, the items in common between two subsequent booklets are given in different order: in the first one they appear at the end and in the second one they appear at the beginning. Comparing the item parameter estimates of two subsequent booklets, we observed that in the second one the items are easier, which could be explained by the position of the items. The effect of the block position is well-known in the literature (see e.g. Rutkowski et al., 2014, p. 236). It should be noted, however, that the TIMSS data are analyzed considering all the forms together, which can mitigate the effect of the item positions.

5 | 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, synthesizing the information 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 problem is detected at the origin. The procedure makes also 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 *SE*s of the equating coefficients, computed as explained in Battauz (2013), 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 conversions.

When the sample size is very large, the random variability is very limited and the test tends to reject 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 4, is still informative about the magnitude of the drift. In this respect, using a criterion such as the difference that matters (Dorans & Feigenbaum, 1994) might help in interpreting the results.

A different approach is the detection of item parameter drift at each step of the chain, considering each pair of forms. For this purpose, statistical tests developed for the detection of differential item functioning (DIF) can be employed (Donoghue & Isham, 1998). However, it should be noted that this is a case of multiple testing, which requires appropriate techniques to prevent the inflation of false positive rates (see e.g. Efron & Hastie, 2016, §15.1). Furthermore, the detection of DIF needs a set of DIF-free items, which is not easy to identify.

It should be stressed that in the IRT framework, the equated scores deriving from different paths can be different if and only if the equating coefficients are different. Hence, any source of error that leads to different equated scores leads also to different equating coefficients. Therefore, the test proposed in this paper is not confined to the detection of shifts of item parameters, but it could reveal differences is the equated scores due to any source of systematic error.

ACKNOWLEDGMENTS

The author would like to thank the editor, the associate editor, and two anonymous reviewers for their constructive comments and suggestions. Open Access Funding provided by Universita degli Studi di Udine within the CRUI-CARE Agreement.

ORCID

Michela Battauz https://orcid.org/0000-0002-3098-689X

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.

Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. *Applied Psychological Measurement*, *11*(3), 279–290.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, *48*(6), 1–29.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, *22*(1), 33–51.

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT*. In N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, I. M. Lawrence, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT*. Princeton, NJ: Educational Testing Service.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science*. Cambridge: Cambridge University Press.

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*(3), 179–193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*(2), 139–160.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3. Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

Mullis, I., & Martin, M. (2013). *TIMSS 2015 assessment frameworks*.

Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement*, *25*(1), 53–67.

Ogasawara, H. (2003). Asymptotic standard errors of IRT observed-score equating methods. *Psychometrika*, *68*(2), 193–211.

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. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rutkowski, L., von Davier, M., & Rutkowski, D. (2014). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton: CRC Press.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Battauz, M. (2023). Testing for differences in chain equating. *Statistica Neerlandica*, *77*(2), 134–145. https://doi.org/10.1111/stan.12277