

# Chapter 9

## Finding Equivalent Standards in Small Samples



Monika Vaheoja

**Abstract** The process of resetting performance standards, with small samples, in different forms of an exam is statistically challenging as the estimates are often biased. Empirical information is therefore, often neglected and content experts reset the standard. In the current article, performance standards are set to a new form in by different methods: circle-arc equating and concurrent calibration with OPLM as an IRT model. The responses on the exam forms that are to be equated are simulated in different situations varying in sample size, test length, test difficulty and respondent's abilities. The results demonstrate that even in small samples (50 subjects taking both tests), the IRT-method with OPLM as a model outperforms circle-arc equating when test difficulty and population ability interact.

### 9.1 Introduction

In computer-administrated tests using item banks, students with different abilities answer different item sets varying in difficulty, discrimination and number of items. When test takers receive a diploma or certificate, such a set of items can be referred to as an exam, through which students demonstrate their mastery of a topic by achieving a certain minimal level. This minimal performance level is reflected in a cut-score which reflects a performance standard for the exam and must be fair: cut-scores set on different exam forms should lead to identical decisions for examinees with the same ability. In other words, the probability to pass a test must be related to the ability of the test taker, not the exam the student has responded to.

Numerous methods are available for setting a performance standard on an exam (for an overview, see Hambleton and Pitoniak 2006) and numerous statistical models are available for test equating (Kolen and Brennan 2004; von Davier 2011). Test equating is a process in which the main goal is to establish, with as near accuracy

---

M. Vaheoja (✉)

University of Twente OMD-BMS, Enschede, The Netherlands

e-mail: [m.vaheoja@utwente.nl](mailto:m.vaheoja@utwente.nl)

10voordeleraar, The Hague, The Netherlands

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, [https://doi.org/10.1007/978-3-030-18480-3\\_9](https://doi.org/10.1007/978-3-030-18480-3_9)

as possible, a valid equivalence between raw scores on multiple forms of a test. For example, to compare examinees' scores on multiple versions or forms (Holland and Rubin 1982). This ensures that test scores from multiple forms can be used interchangeably. However, as the test equating process is a statistical approach, it provides more precise results in larger samples (Kolen and Brennan 2004, p. 307–309). The models and methods that are advised for small samples have been shown to include large equating error and bias (Kim and Livingston 2010). One method that outperformed the advised ones, in the context of small sample testing, was circle-arc equating (Dwyer 2016; Kim and Livingston 2011; LaFlair et al. 2015; Livingston and Kim 2009), an equating method based on classical test theory. This means that circle-arc equating may be preferred for the transfer of performance standards to new forms in cases where there are limited number of examinees.

Dwyer (2016) studied this problem by comparing the estimation precision of the cut-score when content experts set the standard using the Angoff standard-setting approach, to when the cut-score was reset with circle-arc equating. The content experts' estimates were also rescaled by correcting them based on estimates of anchor items. The results showed that circle-arc equating did indeed outperform resetting and rescaling the cut-score in maintaining an equivalent performance standard across exam forms. However, this study had observed score equating as its subject, which introduces equating errors when examinees' ability distributions differ (Lord and Wingersky 1988). When the examinees' abilities differ across forms, item response theory (IRT) is advised.

Item response theory models the probability of a respondent correctly answering an item (Hambleton et al. 1991). A correct score on an item is dependent on the ability of the respondent and on the characteristics of the item. Respondents' ability cannot be observed directly, so it is included in the model as a latent variable ( $\theta$ ). The characteristics of the item can be its difficulty parameter ( $\beta$ ), discrimination parameter ( $\alpha$ ) and guessing parameter ( $\gamma$ ). The primary goal of item response models is to estimate this latent variable for each person as precisely as possible. Limited research is available on the minimum sample size required for IRT equating to maintain equivalent performance standards across test forms. Using IRT equating is even discouraged with small samples (Kolen and Brennan 2004, p. 303).

However, the use of IRT may be recommended for transferring a performance standard from one form to the next in cases where exam form consist of highly discriminative items. Because the  $\alpha$  and  $\beta$  of items influences the test information function and thereby the test standard error (Yen and Fitzpatrick 2006). And if the exam has been constructed to have its average difficulty nearby the targeted expected cut-point, its standard error will be the smallest around that area.

In IRT, the standard error of measurement differs across ability and gives the smallest error in expected score values where the test gives the most information. This means that where the most information is given, the expected scores for the difficulty of the exam are estimated more accurately (Lord and Novick 2008). With the classical test theory approach, which includes circle-arc equating, the standard error of measurement is equal for all scores. If the cut-score falls into the extremes of the scale score, it may be better to use the circle-arc approach due to the constant

standard error across the score scale. However, if the cut-score is in the middle range of the score scale, it may be preferable to use the IRT approach.

Therefore, within this article, both of the above methods are compared in re-estimating the performance standard on a new form in cases where exam forms vary in length, number of examinees, difficulty and ability distribution.

## 9.2 Method

To maintain an equal performance standard across exam forms and to study its accuracy, we used simulated data. Because our interest is to find the best solution for the practice for the national exam in teacher-training program in the Netherlands, we make use of the characteristics of the student ability distributions, characteristics of the item bank and the process to find a cut-score on a new form that equals with the ability of the cut-score on the reference exam. This means that we transferred the cut-score from the reference exam to a new form with IRT concurrent calibration and circle-arc equating.

The examinees' item responses were simulated for two forms to study the effect of sample size (50, 100, 200, 400 and 800), ability (both populations have equal ability, second population has lower ability, and second population has a higher ability), test length (25, 50 and 100 items), and difficulty (both forms are same difficulty, second form is easier and second form is harder) creating a total of  $(5 \times 3 \times 3 \times 3 =)$  135 facets. The data structure is from computer administrated exams for a maths teacher-training program in a public secondary school in the Netherlands.

**Sample size and ability.** For sample sizes of 50, 100, 200, 400 and 800 subjects, item responses were simulated for both forms. The ability distribution of the reference population was equal to the mean of the examinees in the maths teacher-trainees population. By subtracting and then adding 0.4 standard deviation to the average population ability, we created populations with both lower and higher abilities (see Table 9.1). This to make sure we created a similar context as it is in the practice.

**The exams.** Twenty five items were randomly sampled from an item pool that was stratified based on discrimination parameters of the calibrated items. These 25 items defined the reference exam, for which we set a cut-score of 13.456, comparable with the average percentage of the cut-scores that experts had set for all maths exams. The theoretical cut-score on the second form was computed by estimating the expected score that is equal to the ability level on the reference exam.

**Table 9.1** Mean and standard deviation of different ability distributions

	Population ability		
	Lower	Reference	Higher
Mean ability	-0.071	0.139	0.349
Standard deviation of ability	0.525	0.525	0.525

**Table 9.2** Exam difficulties and corresponding cut-points

	Exam		
	Easier	Reference	Difficult
Mean difficulty	-0.071	0.139	0.349
Mean of anchor items	0.274	0.274	0.274
Mean of other items	-0.157	0.105	0.368
Standard deviation of difficulty	0.437	0.406	0.402
Cut-point	16.530	13.456	9.780

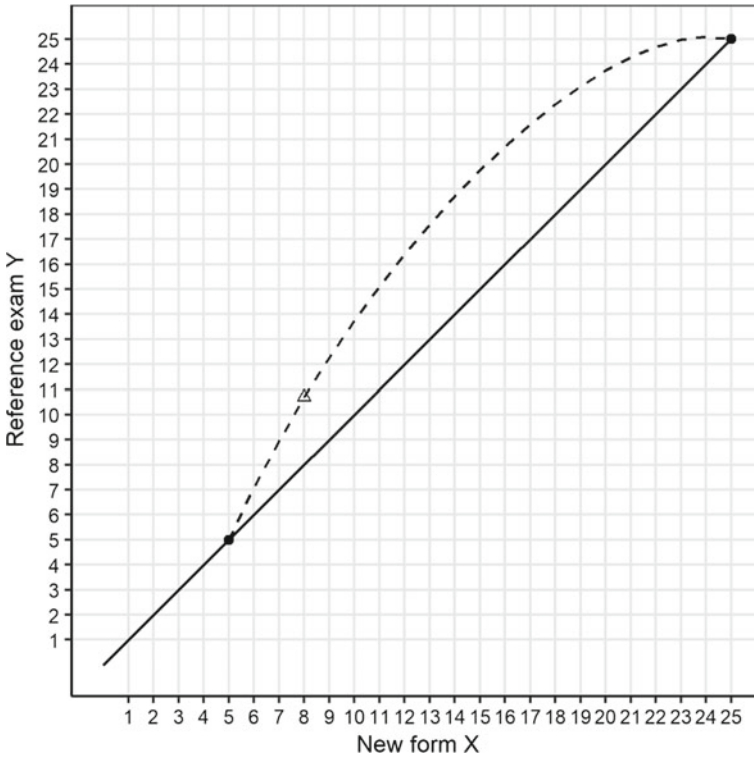
**Test length and difficulty.** To lengthen the test, these 25 items from the reference exam were doubled for the exam with 50 items and quadrupled to create a form with 100 items, which kept all other test characteristics constant. Five items from the 25 were marked as anchor items, anchoring both forms by 20%. The difficulty of the new form was adjusted by subtracting or adding a constant to the difficulty parameters of non-anchor items. The reasoning behind this being that the difficulty of the total exam should be equal to the mean ability of the different populations. This resulted in both easier and difficult forms (Table 9.2).

**General Method.** To maintain an equal performance standard across exam forms with the simulated data, we transferred the cut-score from the reference exam to a new form with IRT concurrent calibration and circle-arc equating. This procedure of simulating response data and transferring the cut-score to a new form in each facet was repeated until we had 1000 successful runs. When the IRT model could not be estimated in a run, because the entire examinee cohort had simulated correct or wrong scores, the circle-arc estimators were excluded, and new data was simulated.

**Circle-Arc equation.** In circle-arc equating, the observed score scales from both forms are equated by assuming a curve-linear relationship between both forms from an arc (see an example in Fig. 9.1; Livingston and Kim 2009). This arc represents an equating function through which the mutually corresponding observed scores on both forms can be related. An arc is mathematically derived and has three points: minimum and maximum possible scores and an empirically estimated midpoint.

In Fig. 9.1, this equating function is illustrated for the new and reference forms. The solid line is the identity line between both forms. In Fig. 9.1, three points are drawn: the lower dot represents the arbitrary chosen minimum possible score, which could be the chance score on both forms. The upper point represents the maximum possible score on both forms. The triangle represents the empirically estimated midpoint, which in the case of non-equivalent anchor test design (NEAT design) is derived from chained linear equating (see Livingston and Kim 2009). If there is no difference between the two forms, then the circle-arc function will be equal to the identity line. If the curve falls to the left, then the reference form is easier. Finally, if the curve falls to the right of the identity line, then the new form is easier.

**Concurrent calibration and OPLM.** In IRT concurrent calibration, exams that have a common set of items are calibrated jointly (Kolen and Brennan 2004). Then, the



**Fig. 9.1** The circle-arc equating function

items are combined into a test function that relates the expected scores on one exam to the ability scale. The test function is the sum of the items from that particular exam. The test function of the second exam is computed in the same manner, which makes it possible to find the expected scores for both exams using a specific ability that is related to the performance standard.

In Fig. 9.2, this process is visualized. The dashed line represents the test function for the reference exam, whilst the solid line represents the new form. In order to find an equivalent cut-score on the second exam, firstly, an ability score that corresponds to the cut-score on the reference exam must be found. In the example in Fig. 9.2, the corresponding ability value equals 0.03. As both forms are calibrated jointly, the same ability score can be used to find an equivalent expected score on the second form. In the lower part of Fig. 9.2, the corresponding expected score is equal to 9.98. From this we can conclude that the second form is more difficult than the reference form as the equivalent cut-score is lower.

The IRT model in the current article is the one parameter logistic model (OPLM), in which elegant features of the Rasch and the Two Parameter Logistic (2PL) models are combined (Verhelst and Glas 1995; Verhelst et al. 1993). The discrimination

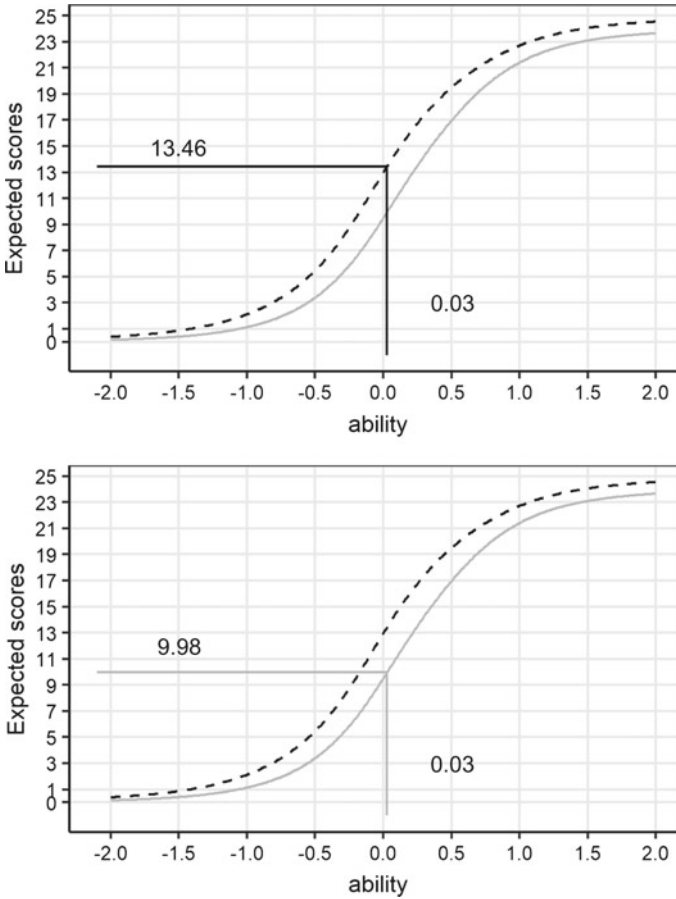


Fig. 9.2 Illustration of IRT concurrent calibration equating to find an equivalent score

parameters ( $\alpha_i$ ) are imputed into the OPLM as a constant ( $a_i$ ). These constants are supplied through a hypothesis. This means that in the OPLM, the items discriminate differently, and the difficulty parameters can be estimated by the conditional maximum likelihood.

### 9.3 Results

Figures 9.3 and 9.4, present the bias and Root Mean Squared Error (RMSE) for each of the facets, per equating method.

**Bias and variability in estimators.** Bias is the most important measure in our case as when the equating method re-estimates the cut-score and it does not correspond

to the same ability, this could have major consequences for the examinees. The corrected bias means that the bias was divided by test length. As there is higher bias for longer tests, more examinees are affected in shorter tests than in longer tests. For example, if forms of 25 and 50 items are administered to 100 students each, there are relatively more students falling in one score range in shorter exam than for a longer exam. Which means that one score difference in shorter exam affects relatively more students than in a longer exam.

Figure 9.3 shows the average corrected bias for each facet in both methods. A negative bias means that the exam was estimated to be more difficult and a positive bias means that the exam was estimated to be easier. The results demonstrate that the IRT estimators are less biased in each facet than the estimators from the circle-arc method. Therefore, it seems as though the circle-arc method tends to estimate a lower cut-score for an easier exam and a much higher one for a more difficult exam, even in contexts where the participants have equal ability. Nevertheless, the IRT method shows some bias too. The bias in IRT estimators decreases with an increased sample size within, test length and ability change. With circle-arc equating, no difference in bias for sample sizes was found, but bias differed across population ability. Circle-arc estimators have less bias for easier exams in cases where the ability of the second population is higher. Additionally, the estimators show less bias for more difficult exams in cases where the population ability is lower.

In general, the estimators of cut-scores in the new forms vary less with IRT (Fig. 9.4). Only in cases where exams have the same difficulty, circle-arc estimators

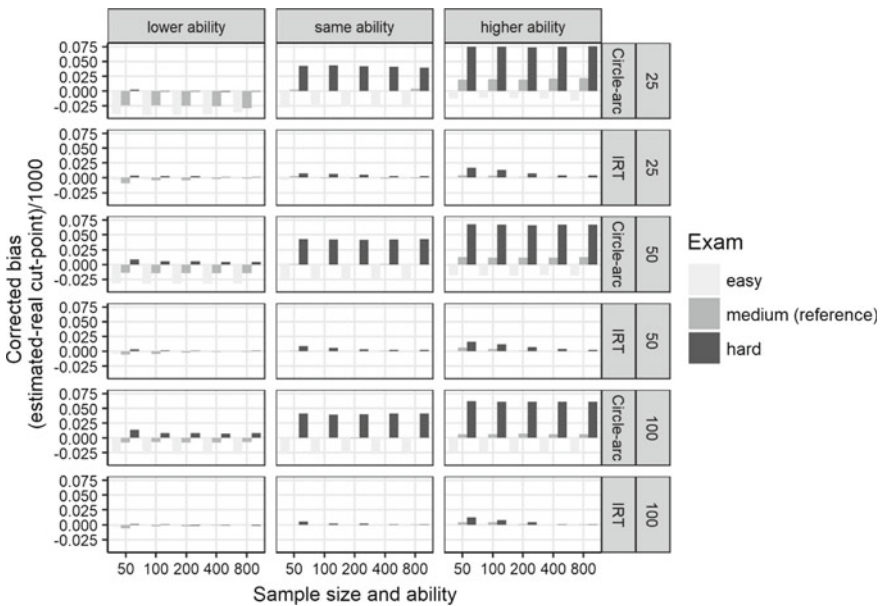


Fig. 9.3 Corrected bias in the estimators from circle-arc equating and IRT

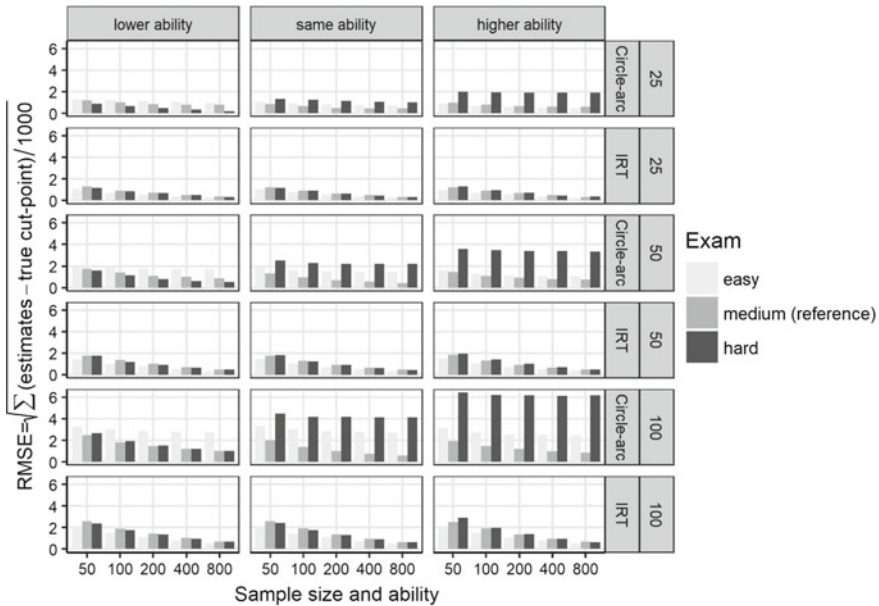


Fig. 9.4 RMSE of the estimators from circle-arc equating and IRT equating

vary less than IRT estimators. The IRT estimators show greater variability for smaller samples, whereas with circle-arc equating, no variability can be seen for different sample sizes. The variability increases with the test length for both methods. Furthermore, the estimators from both methods vary more in estimating the cut-score for difficult exams in contexts where the ability of the examinees is the same or higher. However, in contexts where the examinees from the second population have a lower ability, circle-arc equating varies more in estimating the cut-score for easier exams.

### 9.4 Conclusion and Discussion

The results generally indicate that IRT estimators were less biased and showed less variability in estimating the cut-score in the new form than the estimators of circle-arc equating. Only in contexts where the examinees had the same ability and both forms had the same difficulty showed the circle-arc estimators less bias and variability. However, in such cases, no equating is needed as both forms are parallel. For both methods, we saw an increase in bias for longer tests. Only for IRT estimators we observed a decrease in corrected bias and variability for large samples. Even for small samples of 50 examinees per form, the IRT estimators were less biased than the circle-arc estimators were.



There are two possible reasons for the success of our study in favour of IRT estimators. Firstly, the stratified random items from the item bank were highly discriminating and difficulty of the items was diverse. Secondly, we used OPLM as our IRT model to calibrate both forms and our focus was only on estimating the cut-score. Cut-scores often fall in the middle range of the score scale where the expected test scores have the smallest standard error.

The use of OPLM to calibrate the exams made it possible to use conditional maximum likelihood to estimate the parameters. This is an advantage when the parameters are estimated with a limited number of examinees per form and the examinees were not a random sample from the population. Their average ability, therefore, does not represent the population and it would be ambitious to assume the ability distribution to be normal.

However, some bias was found in the IRT estimators, particularly within small samples. The bias in the estimators was higher in contexts where the examinees from the second form were in the higher ability group and took the more difficult exam. The estimates were higher, meaning that the difficult exam was estimated as easier than it was supposed to be. This bias might have been caused by the anchor set in our study, because the anchor set was more difficult than the mean difficulty of the reference exam. Which is not advised for equating, in fact, the anchor set should be a miniature version of the exam including its difficulty (Holland and Dorans 2006). This, however, might indicate an important exam construction rule, which seems to be more crucial for small samples than for exams with more examinees.

Investigating the impact of anchor set when transferring the performance standard in small samples, where there is no bias in estimators and the anchor set is a miniature of the exam, could be extremely relevant. Another suggestion would be to use a fixed parameter calibration in which the anchor set parameters, or the parameters of the reference exam items are fixed. Kolen and Brennan (2004; p. 183) briefly addressed this aspect and implied that fixing the parameters in contexts where the ability of the populations differs, might lead to biased estimates. This is because the ability distribution is estimated as a mean of zero and a standard deviation of one. However, this is only the case when the marginal maximum likelihood is used to estimate the parameters, this may not be the case when the conditional maximum likelihood is used, as it is in OPLM.

The bias present in the circle-arc method should not be neglected. Even though, Dwyer (2016) showed promising results in favour of the circle-arc method, this method has some weaknesses. Livingston and Kim (2009) present this equating method as an improvement in the chained linear equating method, as a method to overcome impossible maximum scores because it follows an arc. However, in the case of NEAT design, is the empirically estimated midpoint for the circle-arc method derived from the chained linear equating method. If the new form is difficult and examinees abilities are low, this arc still results in impossible maximum scores. The circle-arc method could be improved by adding a second empirical point which partially restricts the arc. A second weakness of the circle-arc equating method is the definition of the minimum score. Authors tend to leave the decision for the minimum

score up to the user. However, the choice of the minimum score affects the length of the circle's radius that is then used to compute the arc and the equating function.

Finally, Hambleton and Jones (1993) observed a limitation of the classical test theory approach; the test scores obtained by classical test theory applications are dependent on the test and biased when examinees' abilities differ. Within this article, this limitation was empirically demonstrated. Additionally, although Kolen and Brennan (2004) did not advise using OPLM as an IRT equating model for small samples, we would urge researchers to consider using OPLM in resetting performance standards due to our promising results.

## References

- Dwyer, A. C. (2016). Maintaining equivalent cut scores for small sample test forms. *Journal of Educational Measurement, 53*, 3–22.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement*. American Council on Education and Praeger Publishers.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement*. American Council on Education and Praeger Publishers.
- Holland, P. W., & Rubin, B. P. (1982). *Test equating*. New York: Academic Press.
- Kim, S., & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement, 47*(3), 286–298.
- Kolen, M. J., & Brennan, R. J. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.
- LaFlair, G. T., Isbell, D., May, L. D. N., Arvizu, M. N. G., & Jamieson, J. (2015). Equating in small-scale language testing programs. *Language Testing, 12*(23), 127–144.
- Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46*, 330–343.
- Lord, F. M., & Novic, M. R. (2008). *Statistical theories of mental test scores*. IAP.
- Lord, F. M., & Wingersky, M. S. (1988). Comparison of IRT true-score and equipercentile observed-score 'equatings'. *Applied Psychological Measurement, 8*(4), 453–461.
- Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models*. New York, NY: Springer.
- Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1993). *OPLM: One parameter logistic model [computer software manual]*. Arnhem.
- von Davier, A. A. (2011). *Statistical models for test equating*. New York, NY: Springer.
- Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement*. American Council on Education and Praeger Publishers.

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

