Modeling the effect of differential motivation on linking educational tests

Keizer-Mittelhaëuser, M.-A.

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Keizer-Mittelhaëuser, M-A. (2014). *Modeling the effect of differential motivation on linking educational tests*. [s.n.].

Modeling the Effect of Differential Motivation on Linking Educational Tests

Marie-Anne Keizer-Mittelhaëuser

Modeling the Effect of Differential Motivation on Linking Educational Tests

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University
op gezag van de rector magnificus, prof. dr. Ph. Eijlander,
in het openbaar te verdedigen ten overstaan van
een door het college voor promoties aangewezen commissie
in de aula van de Universiteit

op vrijdag 12 december 2014 om 14:15 uur

door

Marie-Anne Keizer-Mittelhaëuser,
geboren op 19 augustus 1986 te Lelystad

Promotor:	Prof. dr. K. Sijtsma
Copromotor:	Dr. A. A. Béguin
Overige leden van de Promotiecommissie:	Dr. M. von Davier
	Prof. dr. ir. T. J. H. M. Eggen
	Prof. dr. R. R. Meijer
	Prof. dr. J. K. Vermunt
	Prof. dr. M. P. C. van der Werf

Contents

Chapter 1

Introduction[*]

In educational measurement, multiple test forms are often constructed to measure the same construct to prevent item disclosure and maintain fairness. To make accurate comparisons of results, different test forms are created with as equal content and psychometric properties as possible. However, it is unlikely the test forms will be perfectly comparable. Therefore, score differences between test forms can be attributed either to differences in difficulty of the test forms or to differences in proficiency of the examinees. Equating and linking procedures can be used to disentangle differences between test form difficulty and differences between the proficiency of examinees (von Davier, 2013) so that scores on different test forms can be used interchangeably (see Angoff, 1971; Holland & Rubin, 1982; Kolen & Brennan, 2004). Multiple data collection designs can be considered for collecting data to be used for linking. Choosing one type of data collection design over another depends on practical and statistical limitations. For example, differential student motivation for test taking needs to be considered when choosing a data collection design (Holland & Wightman, 1982). Differential motivation refers to the difference with respect to test-taking motivation that exists between high-

[*] This chapter will be published as: Mittelhaëuser, M., Béguin, A. A., & Sijtsma, K. (in press). Selecting a data collection design for linking in educational measurement: Taking differential motivation into account. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & W.-C. Wang (Eds.), *New Developments in quantitative psychology: Presentations from the 78th Annual Psychometric Society Meeting*. New York: Springer.

stakes and low-stakes administration conditions. In a high-stakes administration condition, an examinee is expected to work harder and strive for maximum performance, whereas a low-stakes administration condition elicits typical, rather than maximum, performance. Even though essentially all data collection designs are effective when all examinees are sufficiently motivated, the way in which data collection designs are typically implemented in practice results in some data collection designs being more robust against the effect of differential motivation than others.

In this introduction, we first discuss differential motivation, followed by an overview and discussion of the robustness of linking procedures against the effect of differential motivation for five well-known types of data collection designs. Then, an example is used to highlight the need to consider differential motivation when choosing a data collection design for linking.

Differential Motivation

Researchers often implicitly assume that a test score is a valid indicator of an examinee's best effort (Wolf & Smith, 1995). However, accumulated evidence shows that if item performance does not contribute to the test score or if no feedback is provided, examinees may not give their best effort (Kiplinger & Linn, 1996; O'Neill, Sugrue, & Baker, 1996; Wise & DeMars, 2005). Unusual patterns of item scores or under-performance are common for low-stakes administration conditions. Within the item response theory (IRT) framework, unusual item-score patterns and under-performance threaten the correct estimation of examinee proficiency and item parameters (Béguin & Maan, 2007). For example, Mittelhaëuser, Béguin, and Sijtsma (2011) found that, compared to using common items administered in a high-stakes condition, using common items administered in a low-stakes condition to link two high-stakes tests yielded different conclusions about the proficiency distributions.

Many studies have focused on preventing, detecting, or correcting the effect of differential motivation. For example, Wise and Kong (2005) pointed out that the effort an examinee devotes to an item may vary throughout the test. Furthermore, Wolf, Smith, and Birnbaum (1995) found that the effect of the administration condition on test performance differs substantially for different groups of items. In particular, items scoring highly on perceived difficulty or items considered mentally taxing were more affected by a difference in administration condition. Despite the growing knowledge of differential motivation, in practice, the effect differential motivation has on data is hard to detect and correct. Reise and Flannery (1996) address this problem by stating, "Typical performance tests are usually not taken as seriously by examinees as are maximum performance measures. … which is potentially more damaging to the measurement enterprise than any of the other so-called 'response biases'" (ibid., p. 12). Since differential motivation might threaten the correct estimation of examinee proficiency and item parameters, thereby threatening the link between two test forms, differential motivation has to be taken into account when choosing a data collection design for linking.

Data Collection Designs

This section provides an overview of five well-known types of data collection designs suitable for linking and addresses the robustness of linking procedures and the way data collection designs are typically implemented in practice against the effect of differential motivation. A detailed description of these data collection designs and a discussion of the general advantages and disadvantages can be found in equating literature (see, e.g., Béguin, 2000; Kolen & Brennan, 2004; Scheerens, Glas, & Thomas, 2003; Von Davier, Holland & Thayer, 2004). A distinction is made between data collection designs in which the tests to be linked are administered to equivalent groups (i.e., single-group design or equivalent-groups design) or to non-equivalent groups (i.e., common-item non-equivalent groups design, pre-test design or linking-groups design). Symbolic representations of the data collection designs are presented in Figure 1.1 through Figure 1.5 in the form of person-by-item matrices. Rows correspond to examinee data and columns to item data. Shaded areas represent combinations of items and examinees for which data are available. Blank areas represent combinations of items and examinees for which no data are available. The ordering of the items presented in the figures does not necessarily correspond to the ordering of items in the test form. Furthermore, sample sizes are not proportional to the sizes of the shaded and blank areas in the figures.

Single-group or equivalent-groups designs

The first data collection design is the single-group design (Figure 1.1). Both test forms are presented to a single group of examinees. An important assumption is that the proficiency of examinees does not change from one test form to the next. By assuming that the proficiency of examinees does not change, score

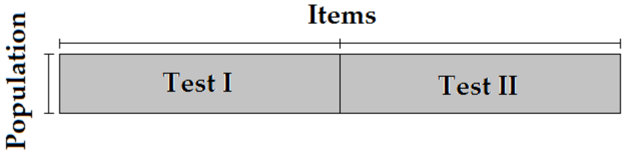


Figure 1.1 *the single-group design*

differences between the two test forms can be attributed to differences in test form difficulty. Differential motivation should not pose a problem when using this data collection design if both test forms are administered under the same (high-stakes) conditions. However, if Test I is administered in a condition where the stakes are higher than in the administration condition of Test II, score differences between the test forms, due to differences in administration conditions, will be attributed to differences in test difficulty, resulting in overestimation of the difficulty of Test II.

The equivalent-groups design (Figure 1.2) is a variation on the single-group design in which each test form is administered to separate, non-overlapping groups of examinees. An important assumption is that the groups are randomly equivalent. By assuming that the groups are randomly equivalent, score differences between the two test forms can be attributed to differences in test form difficulty. Similar to the single-group design, differential motivation should not pose a problem if both tests are administered under the same (high-stakes) conditions. However, if Test I is administered in a condition where the stakes are higher than in the administration condition of Test II, overestimation of the difficulty of Test II is likely.

Kolen and Brennan (2004, pp. 17-19) give an empirical example of differential motivation in a (supposedly, counterbalanced) single-group design. They describe how a dataset collected according to a single-group design was used to scale an old test form and a new test form of the Armed Services Vocational Aptitude Battery (ASVAB) (Maier, 1993). It appeared that many examinees were able to distinguish the items of the old test form and the new test form. Furthermore, many examinees were aware that only the items of the old test form were used to determine the score that was employed for selection purposes. Therefore, examinees were more motivated to answer the items of the old test form than the items of the new test form. This difference in stakes between the items from the old test form and items from the new test form

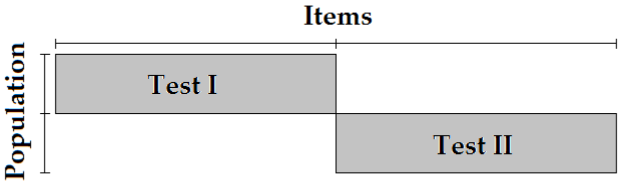


Figure 1.2 *the equivalent groups design*

resulted in high scores on the new test form, resulting in an estimated 350,000 individuals entering the military between January 1, 1976 and September 30, 1980 who should have been judged ineligible.

Non-equivalent groups designs

In non-equivalent groups designs, examinees taking different test forms are assumed to be drawn from different populations. These designs are especially useful when it is unrealistic to assume random equivalence of examinee groups. For example, in educational measurement, the proficiency level of examinee groups may differ. Data in non-equivalent groups designs are collected from the administration of two non-overlapping test forms to two different groups. The data contain no information to disentangle the differences in test form difficulty and the differences in examinees' proficiency. Therefore, non-equivalent groups designs must be 'linked'. Using the common-item non-equivalent groups design, pre-test design or linking-groups design will establish the link in three different ways.

The common-item non-equivalent groups design (Figure 1.3) is the most frequently used data collection design for equating test results across programs and testing organizations (von Davier, 2013). In this data collection design, test forms are administered to non-overlapping and non-equivalent groups of examinees. Both groups, or samples of both groups, are administered an additional set of common items, which are often referred to as anchor items. Since the anchor items are the same across different groups of examinees, the difference in difficulty between the two test forms can be identified from the relative performance of both groups on the anchor items. The common-item non-equivalent groups design has two variations, one using an internal anchor and the other using an external anchor (Kolen & Brennan, 2004, p. 19). When using an internal anchor, the score on the anchor items counts towards the score on the test form, whereas using an external anchor, the score on the

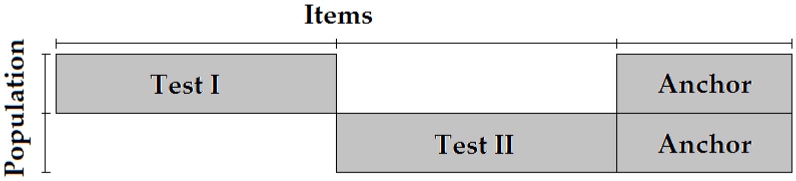


Figure 1.3 *the common-item non-equivalent groups design*

anchor items does not count towards the score on the test form. In an internal-anchor design, the test form and anchor items are administered under the same (high-stakes) administration conditions, and differential motivation should not pose a problem when using this data collection design. Whether differential motivation poses a problem to the external-anchor design depends on the way the design is implemented in practice.

First, differential motivation might be a problem when using an external anchor design if examinees can distinguish which items count towards the score on the test form (i.e., items belonging to the test form) and which items do not (i.e., items belonging to the external anchor). If external anchor items are administered as a separately timed test section, examinees are most likely aware that the scores on these items do not count towards their score on the test form and differential motivation is likely to have an effect. However, if external anchor items are administered at the same time as the test form and examinees are not able to distinguish which items count towards the score on the test form, differential motivation will most likely not pose a problem. Second, differential motivation might be a problem when its effects are unequal between the two populations that are administered the external anchor items. If the effects are equal, differential motivation does not pose a problem and the linking result is unbiased. To see this, one may notice the following. In the common-item non-equivalent groups design the difference in difficulty between the test forms is estimated in two steps. First, the difference in proficiency between the populations is estimated from the relative performance of both populations on the anchor items. Second, the difference in difficulty of the forms is determined based on the relation between the anchor items and the items of the test forms. If the effect of differential motivation is equal among the populations administered the external anchor items, the difficulty of the external anchor items is overestimated, but the relative

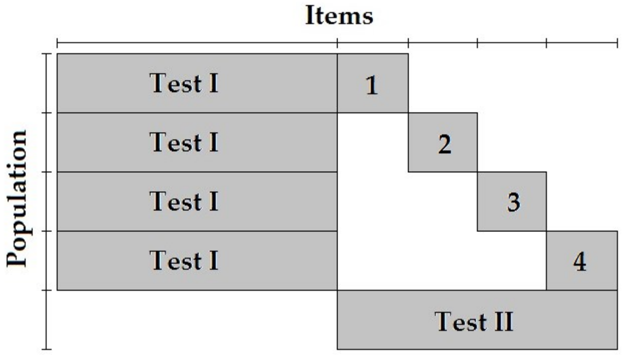


Figure 1.4 *the pre-test design*

performance of both populations on the external anchor items represents the true difference between population proficiency; hence, the linking result is unbiased.

In the pre-test design (Figure 1.4), different subgroups are administered one of the test forms (Test I), and each subgroup receives a different additional subset of items intended for use in the new test form (Test II). In this way, items can be pre-tested to examine their psychometric properties before including them in a test form, here Test II. The score on the pretest items usually does not count towards the score on the test form, since their psychometric properties are unknown at the time of administration. The number of items administered together with Test I is often relatively small to maintain the security of items in the new form (Béguin, 2000). The pretest items should be administered in such a way that the examinees cannot distinguish between the pretest items and the items of the actual test form. In this case, differential motivation should not have an effect on the linking result. However, examinees might be able to distinguish the items of the actual test form and the pretest items, for example, when the pretest items are administered as a separately timed test section. In this case, differential motivation results in an overestimation of the differences in proficiency between the two test forms.

An application of the pre-test design can be found in the standard-setting procedure for the Swedish Scholastic Aptitude Test (SweSat; Emons, 1998; Scheerens et al., 2003). The additional items do not count towards an examinee's score and examinees are not aware of which items do not belong to

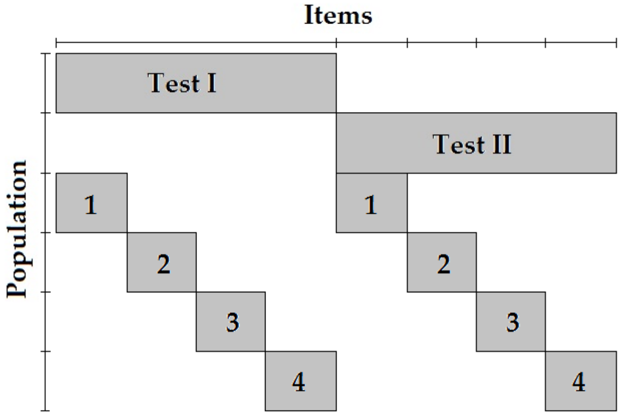


Figure 1.5 *the linking-groups design*

the actual examination, thereby guaranteeing the same level of motivation of the examinees on both the SweSat items and the items that are pre-tested.

Using the linking-groups design (Figure 1.5), a link can be established between the test forms by means of linking groups (Béguin, 2000; Scheerens et al., 2003). Linking groups consists of examinees who do not participate in the actual administration of Test I and Test II but are administered subsets of items from both test forms. Since these examinees are administered subsets of items from both test forms, the difference in difficulty between the two test forms can be estimated from the relative performance of the linking groups on the subsets of items from Test I and Test II. Differential motivation should not pose a problem if the subsets of items are administered to the linking groups in the same (high-stakes) condition as Test I and Test II. If linking groups are administered the items in a lower-stakes condition than Test I and Test II, differential motivation does not necessarily pose a problem. If the effects of differential motivation within the linking groups are equal among the subset of items from Test I and Test II, the linking result is unbiased. To see this, one may notice that if the effects of differential motivation are equal among the subsets of items, the relative performance of the linking groups on the subsets of items from Test I and Test II remains the same and the linking result is unbiased.

Example:
Linking mathematics tests using different data collection designs

This section introduces the mathematics scales of the '*End of Primary School Test*' (Eindtoets Basisonderwijs) and the different data collection designs that can be used for linking the mathematics scales of the End of Primary School Test 2011 and the End of Primary School Test 2012. The linking results obtained using different data collection designs are compared.

End of Primary School Test

The End of Primary School Test is administered each year at the end of Dutch primary education to give students, their parents, and their school advice about the type of secondary education most appropriate for the student. Each year, approximately 80 percent of all primary schools in The Netherlands participate in the test. Even though the advice provided by the End of Primary School Test is not binding, almost all students consider the test high-stakes. This is caused by social- and parental pressure and ample media attention. In addition, some more selective secondary schools use the test scores as part of their admission requirements. Item secrecy is vital; hence, the test form is renewed each year. The test forms of 2011 and 2012 each contained 60 mathematics items.

Method

Data. Samples of students were used to link the mathematics scales. The samples contained 4841 students for the 2011 test form and 5150 students for the 2012 test form.

Data were available to establish the link between the mathematics scales using either an equivalent-groups design (Figure 1.2), a common-item non-equivalent groups design (Figure 1.3) with either an internal or external anchor, a pre-test design (Figure 1.4) or a linking-groups design (Figure 1.5). When using the equivalent-groups design to link the mathematics scales, it was assumed that the samples of 2011 and 2012 were randomly equivalent when estimating the item parameters. Therefore, the differences between the proficiency distributions of the 2011 and 2012 samples did not have to be estimated.

The common-item non-equivalent groups design could be applied to the mathematics scales in two ways, since both an internal anchor and an external anchor were available. When using internal anchor items, samples of students were administered a different test form, which in both cases included 20 anchor items and 40 items from the test form. The anchor items count towards the final score on the End of Primary School Test and students were not aware that they had been presented an alternative test form. Therefore, differential motivation was not expected to pose a problem. The internal anchor items were administered to 3027 and 2708 students in 2011 and 2012, respectively. The external anchor items were administered in a low-stakes condition as a separately timed test. Schools often use this set-up as an additional measurement of proficiency in preparation for the End of Primary School Test. The external anchor test was administered in the same month as the End of Primary School Test. The external anchor test, consisting of 50 mathematics items, was administered to 1696 and 1756 examinees in 2011 and 2012, respectively.

To pre-test the mathematics items intended for use in the End of Primary School Test 2012, 22 pre-test booklets (ranging from 28 to 62 items) were administered in 2011 approximately two to three weeks before the administration of the End of Primary School Test 2011. The number of students who were administered the pre-test booklets ranged from 244 to 347. Since the same pre-test items were administered in more than one pre-test booklet, the number of observations per item was larger, ranging from 276 to 976. The pre-test booklets were administered in a low-stakes condition. Similar to the common-item non-equivalent groups design, the link was established for the 2011 and 2012 samples.

Subsets of items intended for use in the End of Primary School Test 2011 or the End of Primary School Test 2012 were pre-tested on different samples of students to examine the psychometric properties of the items. These samples of students could be used as linking groups in a linking-groups design. Twenty pre-test booklets (ranging from 27 to 63 items) were administered in 2010 approximately two to three weeks before the administration of the End of Primary School Test 2010. The number of students who were administered the pre-test booklets ranged from 150 to 349. Since the same pre-test items were administered in more than one pre-test booklet, the number of observations per item was larger and ranged from 194 to 692. The pre-test booklets were administered in a low-stakes condition.

Analyses. Marginal maximum likelihood estimates of the proficiency distributions of the students who were administered the 2011 or 2012 test forms were obtained using the Rasch model (Rasch, 1960). According to the Rasch model, the probability of passing an item i for student j is a function of proficiency parameter θ_j and can be given by

$$P\left(X_{ij} = 1 \middle| \theta_j\right) = \frac{\exp(\theta_j - \beta_i)}{1 + \exp(\theta_j - \beta_i)},$$

where β_i is the difficulty parameter of item i. OPLM software was used to estimate the Rasch model (Verhelst, Glas, & Verstralen, 1995). The differences in mean proficiency of the 2011 and 2012 samples were compared between the different data collection designs used. Students' *t*-tests were used to determine whether mean proficiency of the samples of 2011 and 2012 differed significantly. Cohen's *d* was used to assess the effect size (Cohen, 1988).

It may be argued that the Rasch model properties of unidimensionality, nonintersecting response curves, and a zero lower asymptote may not be appropriate for the data sets investigated here. However, Béguin (2000) showed that the procedure involving the Rasch model for equating the examinations in the Netherlands is robust against violations of unidimensionality and guessing. We assumed that this result could be generalized to our data and that the use of the Rasch model was appropriate. To investigate whether this assumption wasvalid, the data analysis was repeated on item sets from which items that did not fit the Rasch model were removed.

Results

Table 1.1 shows the estimated proficiency means of the mathematics scales of the End of Primary School Test 2011 and 2012. For all data collection designs, the mean proficiency of the population presented with the 2012 test form was higher than the population presented with the 2011 test form. All effects were significant at a .01 level, but the effect size is considered to be very small when using the common-item non-equivalent groups designs or the linking-groups design, and medium when using the pre-test design (Cohen, 1988). It appears as if differential motivation has a noticeable effect on the resulting link when using a pre-test design with link items administered in a low-stakes condition.

Item misfit was investigated using Infit and Outfit statistics (Wright & Masters, 1982) available in the eRm package in R (Mair, Hatzinger, & Maier, 2010). In scale construction, items having an Infit Mean Square value or Outfit Mean Square value outside the range of 0.5–1.5 (Linacre, 2002) are usually not selected. Items of the End of Primary School Test and the internal anchor had Outfit Mean Square and Infit Mean Square statistics between 0.5 and 1.5, indicating that the Rasch model was consistent with these items (Linacre, 2002). Among the external anchor items, one item had an Outfit Mean Square statistic of 2.031. From the 467 items, which were pre-tested in 2011 and used to link the test forms according to a pre-test design, 14 items had an Outfit Mean Square

Table 1.1 *Estimated proficiency distributions using different data collection designs*

Data collection design	Population	*N*	*M*	*SD*	Cohen's *d*/ Sign. Student's *t*
Common-item internal	2011	4,841	1.232	1.038	0.07 / **
	2012	5,150	1.306	1.064	
Common-item external	2011	4,841	1.133	1.037	0.07 / **
	2012	5,150	1.208	1.062	
Pre-test design	2011	4,841	0.050	1.036	0.47 / **
	2012	5,150	0.547	1.061	
Linking-groups design	2011	4,841	1.176	1.037	0.12 / **
	2012	5,150	1.303	1.062	

** $p < .01$

statistic higher than 1.5. A total of 516 items were pre-tested in 2010 and used to link the test forms according to a linking-groups design, of which 15 items had an Outfit Mean Square statistic higher than 1.5. Given the total number of items, the small numbers of misfitting items indicate that the Rasch model is consistent with these datasets. Deleting the misfitting items in the different data collection designs led to the same conclusion, which is the overestimation of the difference in proficiency distributions when using a pre-test design.

Discussion

Empirical data analyses illustrate the potential effect of differential motivation on results of linking using different data collection designs. Since there is no reason to assume that differential motivation affects the linking result when using a common-item non-equivalent groups design with an internal anchor, the different linking results can be compared with the linking result of this data collection design. The results suggest that the equivalent-groups design is not appropriate for linking both test forms of the End of Primary School Test, since there is a small, although significant difference in proficiency distributions between the samples who were presented either the 2011 or the 2012 test form. Even though students were aware that the items of the external anchor test did not count towards the score on the End of Primary School Test, both common-item non-equivalent groups designs provide the same result. The most likely explanation for this result is that the effects of differential motivation are approximately equal for both populations administered the external anchor test, which leads to the unbiased estimation of the difference between the proficiency of both populations. The same explanation is likely for the linking-groups design, on the basis of which the same conclusion has to be drawn as for both common-item non-equivalent groups designs. Even though all types of data collection designs led to the conclusion that the mean proficiency of the population presented with the 2012 test form was significantly higher compared to the population presented with the 2011 test form, the effect size when using the pre-test design was larger compared to the other data collection designs. Using a pre-test design with linking items administered in a low-stakes administration condition produced differential motivation causing an overestimation of the difference in proficiency distributions, which is consistent with expectation.

All data collection designs may be effective provided all students are sufficiently motivated. However, the way in which data collection designs are typically implemented in practice results in some data collection designs being more robust against the effect of differential motivation than others. The conclusions with respect to the different data collection designs can therefore only be generalized to the extent that data collection designs are implemented in the same way as they were implemented for the End of Primary School Test. To illustrate this, the link items used in the external anchor design, pre-test design and linking-groups design are administered as separately timed tests in low-stakes conditions. The differences between the data collection designs with respect to the estimated proficiency distributions will undoubtedly be negligible if the link items are administered in high-stakes conditions. Furthermore, we expect that administering the link items in a low-stakes condition at the same time as the End of Primary School Test with students being able to distinguish link items and items from the test form, results in larger differences between the data collection design with respect to the estimated proficiency distributions. To see this, one may notice that under these circumstances the difference in performance on the link items and the items from the test form is expected to be larger, since students are likely more inclined to spend effort on answering items correctly from the test form than the link items.

The question that remains is how the effect of differential motivation can be modeled. For example, when items are administered in a low-stakes administration condition, is it possible to classify item-score vectors as either resulting from motivated or unmotivated performance? If this is true, a mixture IRT model with latent classes might be useful for linking high-stakes tests when differential motivation is known to have an effect (Mittelhaëuser, Béguin, & Sijtsma, 2013). Alternatively, examinees might be motivated to a certain degree to answer items correctly in which case a multidimensional IRT model (Embretson & Reise, 2000; Reckase, 2009) might be useful. Furthermore, person-fit methods (e.g., Meijer & Sijtsma, 2001) may be used to investigate how differential motivation affects the individual item-score vector. Since the results suggest that differential motivation has an effect on the linking result in different data collection designs, using methods that produce greater insight into the effect differential motivation has on linking tests administered in a high-stakes condition is valuable for measurement practice and measurement research.

Outline of the thesis

This thesis focuses on modeling the effect of students' differential motivation on linking in educational measurement. I evaluated the performance of the mixture Rasch model using a simulation study and real-data applications. Additionally, the performance of person-fit methods was evaluated using real-data applications. Each chapter in this thesis was written as a research article. Each chapter can be read independently of the other chapters. As a result, the chapters show some overlap, particularly with respect to the introduction of concepts, definitions and notation.

In Chapter 2, I investigated whether the use of the mixture Rasch model helps to diminish the effect of differential motivation on linking two operational versions of the End of Primary School Test. In Chapter 3, I simulated differential motivation between the stakes for operational tests and anchor items and investigated whether linking of the operational tests by means of the Rasch model produces an invalid linking result. Additionally, the performance of the mixture Rasch as a method for modeling simulated differential motivation was evaluated. In Chapter 4, I explored to what extent a mixture Rasch model and the well-known l_z person-fit statistic could be used to model motivational differences in data administered in a low-stakes administration condition. In Chapter 5, I used person-fit methods to investigate the difference between responding in low-stakes and high-stakes administration conditions with respect to test performance and response consistency. In the epilogue, I will reflect on the decisions underlying the operationalization and the modeling of differential motivation.

Chapter 2

Mixture Rasch Modeling of Differential Motivation in IRT Linking[*]

Abstract

The way in which data collection designs used for linking are usually implemented in practice, might make some data collection designs more robust against the effect of differential motivation compared to others. Data from a Dutch testing program were used to investigate whether the differences in estimated proficiency distributions between two operational tests differed between data collection designs with anchor items administered in low-stakes conditions on the one hand and data collection designs with anchor items administered in high-stakes conditions on the other hand. Some data collection designs were found to be more robust against the effect of differential motivation than others. Specifically, the pre-test design resulted in a substantial overestimation of the difference between the estimated mean proficiency of the populations administered the operational tests. The effect of differential motivation in the pre-test design was controlled for by using a mixture Rasch model to link the operational tests. Removing items displaying DIF between high-stakes and low-stakes administration conditions did not improve the linking result.

[*] This chapter has been submitted for publication

Many testing programs use a new test form at each major administration to maintain fairness and prevent item disclosure (Holland & Rubin, 1982). The test forms may differ with respect to difficulty and, as a result, scores on different test versions may not be directly comparable. Several procedures such as linking, scaling, and equating develop a common metric between test forms (e.g., see Kolen & Brennan, 2004). A frequently used data collection design for linking in educational testing is the common-item non-equivalent groups design. In this design, two different test forms are administered in two different populations, for example, sixth-grade primary-school students in two successive years, and both test forms are linked by means of common items (i.e., anchor items). In an educational testing context, the common-item non-equivalent groups design is often eligible, since usually it cannot be assumed that populations are equivalent. For example, the proficiency level of students may change from year to year, producing populations that vary by proficiency level. The common-item non-equivalent groups design can produce a common scale, which enables direct comparison of scores over test forms. The test forms to be linked and the common items are henceforth referred to as operational test forms and the anchor items, respectively.

The validity of the linking result depends on whether the anchor items measure the same construct as the operational tests (Klein & Jarjoura, 1985). In an item response theory (IRT) context, this means that the anchor items and the operational test forms measure the same latent proficiency and are consistent with the same IRT model. The validity of the linking result might be threatened by the effect of differential motivation. Differential motivation refers to the difference in test-taking motivation that exists between high-stakes and low-stakes administration conditions. In a high-stakes administration condition, a student is expected to work harder and strive for maximum performance, whereas a low-stakes administration condition elicits typical, rather than maximum, performance. Empirical evidence for the effect of differential motivation was provided by Wise and DeMars (2005), who found that students might not give their best effort in low-stakes assessment, when they know they receive neither grades nor credit for their performance. In practice, testing programs may use anchor items administered in low-stakes administration conditions to link operational tests administered in high-stakes conditions (Wise & Kong, 2005). This difference between administration conditions may result in unusual patterns of item scores or in relatively meagre performance on the anchor items, and the effect may introduce bias in the linking procedure.

A mixture IRT model may be used to test whether, compared to the high-stakes operational test, the low-stakes condition of the anchor items differently affects some of the item-score vectors (i.e., the vector of item scores a student has produced). Mixture IRT models assume that the data are a mixture of different data sets from two or more latent populations (Rost, 1997; Von Davier & Yamamoto, 2004), also called latent classes, in which populations perform differently on the test items. If this assumption is correct, a particular IRT model does not hold for the entire population, but different model parameters are valid for different subpopulations. Usually, the number of subpopulations and the size of the subpopulations are unknown. In linking high-stakes operational tests with anchor items administered in low-stakes conditions, one can specify the mixture IRT model in such a way that one of the latent classes represents high-stakes responding represented by vectors of item scores unique to this latent class, while the other latent class represents low-stakes responding (Béguin, 2005; Béguin & Maan, 2007). Using solely the data of the latent class representing high-stakes responding in the linking procedure is expected to improve the results of the linking procedure.

Instead of excluding the latent class displaying low-stakes responding from the linking procedure, an alternative procedure to improve the link is to exclude items from the linking procedure, which show differential item functioning (DIF) between low-stakes and high-stakes administration conditions. This procedure is especially useful for investigating whether differential motivation affects particular item types, such as items near the end of a test or relatively difficult items.

We used data from a Dutch testing program to investigate whether the differences in estimated proficiency distributions between two operational tests differ between data collection designs with anchor items administered in low-stakes conditions on the one hand and data collection designs with anchor items administered in high-stakes conditions on the other hand. This was done by comparing the estimated mean proficiency differences of the operational tests over the different data collection designs. Furthermore, the mixture Rasch model was used to investigate whether a latent class representing low-stakes responding could be identified. Next, the latent class that was identified as representing low-stakes responding was removed from the data to diminish the effect of differential motivation on the linking result. Then, item-misfit between high-stakes and low-stakes administration conditions was investigated. Finally, it was investigated whether removing items showing DIF between high-stakes and low-stakes administration conditions diminishes the effect of differential motivation on the linking result.

Method

Participants and Design

Data were used from the mathematics scales of the '*End of Primary School Test*' (Eindtoets Basisonderwijs). This test is administered every year at the end of Dutch primary education, and students' results are used to give advice to schools and parents about the most appropriate type of secondary education. Even though the advice the End of Primary School Test provides is not binding, almost all students consider the test high-stakes because of social and parental pressure and ample media attention. Since the test is administered in a high-stakes condition, item secrecy is vital; hence, the test form is renewed each year. A link between test forms administered in subsequent years can be established using an internal anchor, an external anchor, and pre-test data or a combination of these methods. We developed a common metric for the mathematics scales of the End of Primary School Test using two consecutive test forms, which are the test forms administered in 2009 and 2010, henceforth referred to as 2009 operational test and 2010 operational test, respectively.

Data collection designs. We discuss three data collection designs in which an internal anchor, an external anchor, or pre-test data are used to link the 2009 operational test and the 2010 operational test. The way in which data collection designs used for linking are usually implemented in practice, might make some data collection designs more robust against the effect of differential motivation compared to others. The difference between the designs most relevant to this study concerns whether the anchor items were administered under the same high-stakes conditions as the operational tests, and thus, whether differential motivation can be expected to have an effect on the linking result. Figures 2.1 through 2.3 present symbolic representations of the data collection designs in the form of person-by-item matrices. Rows correspond to student data and columns to item data. The shaded areas represent combinations of items and students for which data are available, while the blank areas represent combinations of items and students for which no data are available. The ordering of the items presented in the figures does not necessarily correspond to the ordering of items in the test form. Furthermore, sample sizes are not necessarily proportional to the sizes of the shaded and blank areas in the figures.

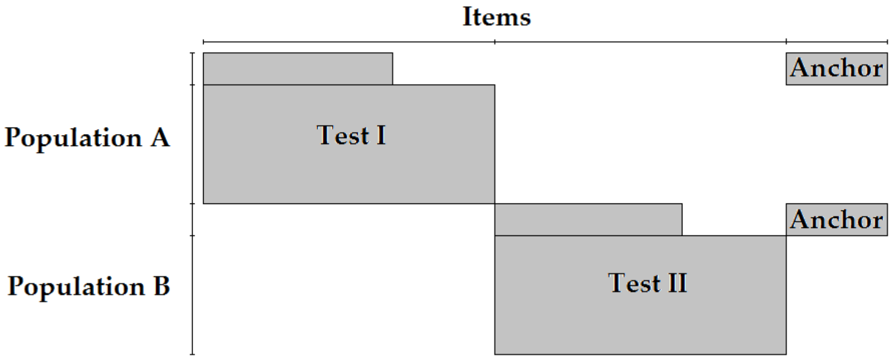


Figure 2.1 *The internal anchor design*

Figure 2.1 shows the internal anchor design. In this design, samples of both populations of students administered the operational test forms were administered an alternative test form, which included a selection of items from the operational test form and additional anchor items, herein referred to as internal anchor items. The anchor items are the same across both alternative test forms. Therefore, differences in difficulty between the operational test forms can be estimated based on the relative performance of the samples on the anchor items. The internal anchor items were placed in the same position in both alternative test forms to avoid undesirable order effects. In this data collection design, all items including the anchor items contribute to a student's score on the operational test. Since the operational tests and the anchor items are administered in the same high-stakes condition, when using an internal anchor design, differential motivation should not have an effect on the linking result. Test security might be threatened if every student who was administered one of the operational tests is presented the anchor items. However, in the internal anchor design, only samples of both populations are presented the anchor items and the students in these samples were not aware that they were presented an alternative test form. The internal anchor items, the number of internal anchor items, and the number of students who were presented the internal anchor items were not made public. It can thus be assumed that the threat to test security is minimal when implementing the internal anchor design in this way. Since the effect of differential motivation is expected to be absent, the internal anchor design is especially useful in linking the 2009 and 2010 operational tests. The difference in estimated proficiency distribution between the two operational tests found for the internal anchor

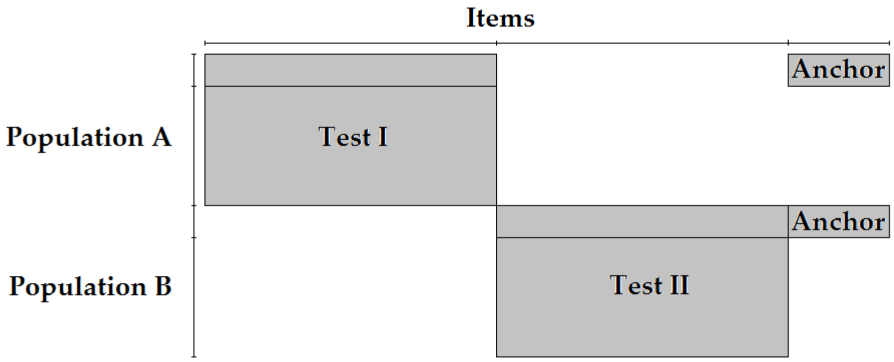


Figure 2.2 *The external anchor design*

design can therefore serve as a benchmark for investigating the effect of differential motivation in different data collection designs.

The external anchor design (Figure 2.2) is different from the internal anchor design in that the score on the anchor items does not contribute to the score on the operational test and the external anchor does not replace part of the operational test. Therefore, when using an external anchor design to link two operational tests, the link between the operational tests is based on the additional anchor items, herein referred to as external anchor items. A distinction can be made between the implementation of this design, where students are either aware or unaware that the score on the external anchor items does not contribute to their score on the operational test, the former being a more serious problem with respect to differential motivation. In applying this data collection design to the End of Primary School Test, students were aware that the score on the external anchor items did not contribute to the score on the End of Primary School Test. As a result, the stakes of the administration condition of the external anchor items and the operational tests differed.

The third data collection design used in this study is the pre-test design (Figure 2.3). Subsets of items intended for use in the operational tests were pre-tested in different samples of students to examine the statistical characteristics of the items before including them in an operational test. Items with the most promising item characteristics were selected for the operational test. However, items that were pre-tested in two consecutive years can be used as an external link between the two operational tests. Similar to the anchor items in the external anchor design, the stakes of the administration condition of the anchor items in

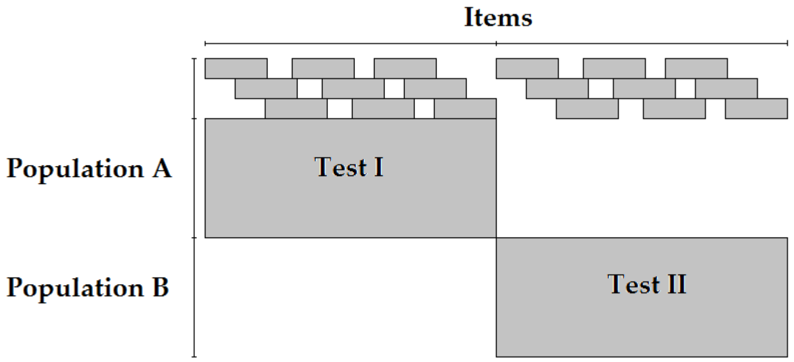


Figure 2.3 *The pre-test design*

the pre-test design differed from the stakes of administration condition of the operational tests.

Participants. Each operational test form contained 60 mathematical items, which did not overlap between the operational test forms. Samples of students administered the operational test forms contained 4,995 students in 2009 and 5,123 students in 2010. The internal anchor consisted of 20 mathematical items administered to 2,989 students in 2009 and 2,421 students in 2010. The external anchor consisted of 20 mathematics items administered to 5,086 students in 2009 and 4,575 students in 2010. In order to pre-test items for the 2009 operational test, in 2008, 19 pre-test booklets (ranging from 30 to 90 mathematical items) were administered. The number of students administered the pre-test booklets ranged from 183 to 313. Since the same pre-test items were administered in more than one pre-test booklet, the numbers of observations per item were larger, ranging from 219 to 1,685. The mathematical items for the 2010 operational test were pre-tested in 2009 using 23 pre-test booklets (ranging from 29 to 60 mathematical items). The number of students who were administered these pre-test booklets ranged from 46 to 372. The number of observations per item ranged from 504 to 1,664.

Analyses

The Rasch model. The Rasch model (Rasch, 1960) was fitted to the data to inspect differences in estimated mean proficiency between the populations administered the 2009 and 2010 operational tests. According to the Rasch model, the probability that student *j* passes item *i* is a function of proficiency parameter θ_j and is given by

$$P(X_{ij} = 1|\theta_j) = \frac{\exp(\theta_j - \beta_i)}{1 + \exp(\theta_j - \beta_i)},$$

where β_i is interpreted as the difficulty parameter of item i. The OPLM software (Verhelst, Glas, & Verstralen, 1995) was used to estimate the Rasch model. The fit of the Rasch model to the items from the operational test was investigated by means of Infit and Outfit statistics (Wright & Masters, 1982) available in the eRm package in R (Mair, Hatzinger, & Maier, 2010). Mean Square Outfit values or Mean Square Infit values above 1 indicate greater variation in the data than predicted by the Rasch model, and values below 1 indicate that the data over-fit the model. Items having statistics outside the range of 0.5 – 1.5 are less useful for Rasch measurement (Linacre, 2002) and were therefore iteratively removed from the analyses. The differences between estimated mean proficiency of both operational tests were compared for each of the three linking designs. Cohen's d was used to assess effect size (Cohen, 1988). Student's t-tests were used to determine whether the differences between the means of the 2009 and 2010 operational tests were significant.

The standard deviations of the estimated proficiency distributions that OPLM provides were used to evaluate the significance of the differences between the estimated mean proficiencies of the populations administered the 2009 and 2010 operational tests. However, instead of using the complete variance-covariance matrix, with larger data sets OPLM only uses the diagonal of the matrix (Verhelst, Glas, & Verstralen, 1995), which might result in an underestimation of the standard deviations of the estimated proficiency distributions. Therefore, a bootstrap procedure (Efron & Tibshirani, 1993) was used to construct 95% confidence intervals (CIs) for the differences between the mean proficiencies of the operational tests. The bootstrap procedure was done using the following sequence of steps:

1. From each dataset obtained from the students administered the internal anchor, the external anchor, or pre-test data, 1,000 bootstrap samples were drawn. Data from the operational tests were not bootstrapped; hence, the same data were used throughout.
2. OPLM was used to estimate the mean proficiency for each population administered an operational test. The analysis was repeated for each bootstrap sample.
3. For each data collection design, steps 1 and 2 resulted in 1,000 differences in estimated mean proficiency between the operational tests. For each type of data collection design, the Shapiro-Wilk test

(Shapiro & Wilk, 1965) was used to test whether the differences found in the bootstrap samples were normally distributed. A 95% CI was constructed using the .025 and .975 percentiles under the normal distribution.

The mixture Rasch model. Next to the Rasch model, the mixture Rasch model was used to link the operational tests using the three different data collection designs. Let X_i denote the score on item *i*, with the total number of items represented by *k*. According to the mixture IRT model, the probability of passing item *i* depends on a class-specific proficiency parameter θ_{jg}, denoting the proficiency of student *j* if he/she belongs to latent class g. The techniques currently available for estimating a mixture IRT model focus on the Rasch model. The limitation to the Rasch model is partly because of the limited information in the data to estimate more-complex models. The mixture Rasch model defines the conditional response probability as

$$P(X_{ij} = 1|\theta_{jg}) = \frac{\exp(\theta_{jg} - \beta_{ig})}{1 + \exp(\theta_{jg} - \beta_{ig})},$$

where β_{ig} is a class-specific difficulty parameter. Aggregated over items, the probability of obtaining an item-score vector $\boldsymbol{x}_j = \{x_{1j}, x_{2j}, \dots, x_{kj}\}$ given proficiency θ_j and membership of class g is

$$P(\boldsymbol{x}_j|g, \theta_j) = \prod_{i=1}^{k} \frac{\exp[x_{ij}(\theta_{jg} - \beta_{ig})]}{1 + \exp(\theta_{jg} - \beta_{ig})}.$$

Let π_g denote the proportion of the population that belongs to class g ($g = 0, \dots, G$). Proportion π_g is also called the class probability. The probability for an individual *j* with item-score vector $\boldsymbol{x}_j$ to belong to class g, denoted $P(g|\boldsymbol{x}_j)$, depends on the item-score vector in the following way:

$$P(g|\boldsymbol{x}_j) = \frac{\pi_g P(\boldsymbol{x}_j|g)}{\sum_{g=0}^{G} \pi_g P(\boldsymbol{x}_j|g)}. \tag{2.1}$$

A dedicated version of the OPLM software (Béguin, 2008) was used to estimate the mixture Rasch model. Two latent classes were specified in the mixture IRT model, the first representing responding expected under high-stakes conditions and the second representing responding expected under low-stakes conditions. Classes were specified by modelling the item-score vectors of the operational tests as being exclusively part of the first latent class, which was done by setting $\pi_1 = 1$ and $\pi_2 = 0$ in Equation 2.1 for all item-score vectors of the operational tests. The constraints should identify the first latent class as students showing high-stakes responding and the second latent class as students showing low-stakes responding. The item-score vectors of the external anchor and the pre-test data could be in either the first or the second latent class. As with the simple Rasch model, a bootstrap procedure was used to construct 95% CIs for the differences between the mean proficiencies of the operational tests. The mixture Rasch model was compared to the simple Rasch model to investigate whether modelling subpopulations would improve the link between the operational tests. This was done by means of (1) comparing the differences in mean proficiency between the samples administered the two operational tests, (2) a test for model comparison, and (3) comparing difficulty parameters estimated for both the Rasch model and the mixture Rasch model.

Differential Item Functioning. Instead of identifying item-score vectors from the data, which were affected by the low-stakes nature of the administration condition of the anchor items, one could also remove items from the dataset that function differently in high-stakes and low-stakes administration conditions. DIF analysis identifies items that display different statistical properties in different group settings after controlling for differences between the estimated proficiencies of the groups (Holland & Wainer, 1993). Items displaying DIF between high-stakes administration conditions and low-stakes administration conditions are not suited for establishing a common metric between the operational test forms. In the context of DIF, OPLM software provides the contribution of each item to the R_{1c} statistic (Glas, 1989), which evaluates the squared difference between expected and observed proportions of item-correct scores in homogeneous score groups (i.e., groups in which each student has the same number-correct score).

Items having a mean contribution to the R_{1c} statistic in excess of 4 were selected for visual inspection of DIF. OPLM provided graphs displaying the item characteristic curves (ICCs) for different groups, which were used for visually inspecting DIF between administration conditions. This DIF approach was applied only to the pre-test design, since this is the only data collection design in which anchor items were administered in both high-stakes and low-stakes administration conditions. After removing items displaying DIF, the Rasch model was again fitted to the data and the differences between estimated mean proficiencies of the operational tests were compared to the linking result for the internal anchor design, serving as a benchmark.

Table 2.1 *Proficiency distributions estimated with the Rasch Model.*

Design	Population	*M*	*SD*	*N*	Cohen's *d*/ *p*-value Student's *t*	95% CI
IA	2009	0.000	1.004	7,984	0.015/.362	(-0.023; 0.053)
	2010	0.015	1.046	7,544		
EA	2009	0.000	1.002	10,081	0.068/<.001	(0.038; 0.102)
	2010	0.070	1.044	9,698		
Pre-test	2009	0.000	0.956	11,789	0.375/<.001	(0.355; 0.403)
	2010	0.379	1.062	5,123		

Note. IA = Internal Anchor; EA = External Anchor.

Results

Rasch Analysis

None of the items from the operational test forms had a Mean Square Outfit statistic or Mean Square Infit statistic outside the (0.5–1.5) range, indicating that the Rasch model was consistent with these items (Linacre, 2002). Table 2.1 shows the estimated proficiency means of the 2009 and 2010 operational tests. For the internal anchor design, the estimated mean proficiency of the population administered the 2010 operational test was not significantly higher than the estimated mean proficiency of the population administered the 2009 operational test ($p > .05$). However, for the external anchor design and the pre-test design the estimated difference in mean proficiency between the populations administered the 2009 and 2010 operational tests was significant. Interestingly, the effect size resulting from using the pre-test design was considerably higher than the effect size resulting from using either the internal or external anchor design.

The 95% CIs for the differences between estimated mean proficiencies of the 2009 and 2010 operational tests are presented in Table 2.1, Column 7. The Shapiro-Wilk test (Shapiro & Wilk, 1965; results not tabulated) indicated that the differences between estimated mean proficiencies were normally distributed ($p > .05$) for each data collection design. Even though the results of the Student's *t*-tests led us to conclude that the linking result differed between the internal and external anchor design, the CIs of both designs overlapped. However, the CI for the pre-test design did not overlap with the CIs for the internal and external anchor design. Therefore, we have reason to conclude

Table 2.2 *Proficiency distributions estimated with the mixture Rasch model.*

Design	Population	*M*	*SD*	*N*	Cohen's *d*/ *p*-value Student's *t*	95% CI
IA	2009	0.000	1.008	7,984	0.016/.332	(-0.026; 0.058)
	2010	0.016	1.049	7,544		
EA	2009	0.000	1.007	10,081	0.068/<.001	(0.026; 0.104)
	2010	0.070	1.049	9,698		
Pre-test	2009	0.000	0.885	11,789	0.008/.613	(-0.055; 0.074)
	2010	-0.008	1.070	5,123		

Note. IA = Internal Anchor; EA = External Anchor.

that the results of the linking procedure differed between the different data collection designs. Specifically, compared to the internal anchor design, the pre-test design results in a substantial overestimation of the difference between proficiency distributions of the populations that were administered the operational tests.

Mixture Rasch model

Table 2.2 shows the estimated proficiency means and the corresponding 95% CIs for the 2009 and 2010 operational tests that resulted from the application of the mixture Rasch model. The results for both the internal and external anchor design resemble the results for the simple Rasch model. For the internal anchor design, the estimated mean proficiency of the population administered the 2010 operational test was not significantly higher than the estimated mean proficiency of the population administered the 2009 operational test ($p > .05$). For the external anchor design, estimated mean proficiency between the populations administered the 2009 and 2010 operational tests differed significantly. The mixture Rasch model and the Rasch model provided different results for the pre-test design. For the Rasch model, we found a large effect size, indicating an overestimation of the difference in mean proficiency. However, this large effect diminishes when using the mixture Rasch model, and the conclusions based on the pre-test design (i.e., no significant difference in proficiency between populations administered the operational tests) resemble the conclusions based on the internal anchor design.

For the internal and external anchors, the class membership probabilities of belonging to the latent class representing high-stakes responding were approximately equal in 2009 and 2010: for the internal anchor, both

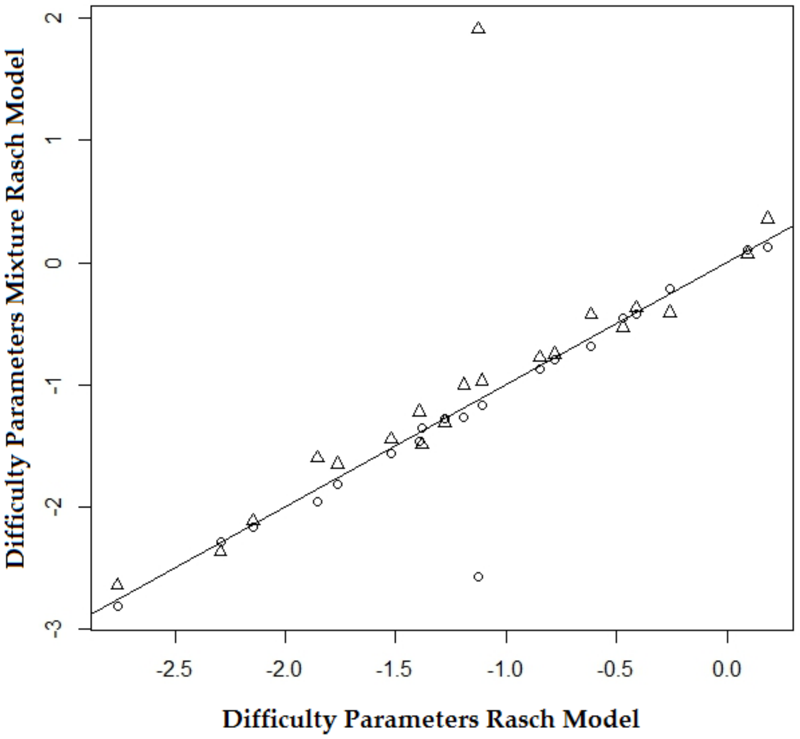


Figure 2.4 *Item difficulty parameters of the Rasch model and the mixture Rasch model for the latent class representing high-stakes responding (○) and the latent class representing low-stakes responding (Δ) estimated in the internal anchor design.*

probabilities were .75 (2009 and 2010), and for the external anchor, they were .62 (2009) and .63 (2010). The class-membership probabilities differed among the pre-test booklets. From the 19 pre-test booklets used to pre-test the items of 2009, one booklet was removed from the analysis because the probability to belong to the class representing high-stakes responding was .00. The mean probability of the remaining 18 pre-test booklets was .58 (ranging from .40 to .79), whereas the mean probability of the 23 mathematics pre-test booklets to pre-test the items of 2010 was .66 (ranging from .30 to .95). The question remains whether the mixture Rasch model provides a better fit to the data than the simple Rasch model and whether the latent classes identified represent low-stakes and high-stakes responding.

Based on the log likelihoods, the mixture Rasch model provided a better fit to the data than the simple Rasch model. This result was found for the internal anchor design (χ^2(138, *N* = 15,528) = 4,158.8, $p < .001$), the external anchor design (χ^2 (138, *N* = 19,779) = 5,800.6, $p < .001$), and the pre-test design (χ^2 (647, *N* = 21,753) = 13,922.8, $p < .001$).

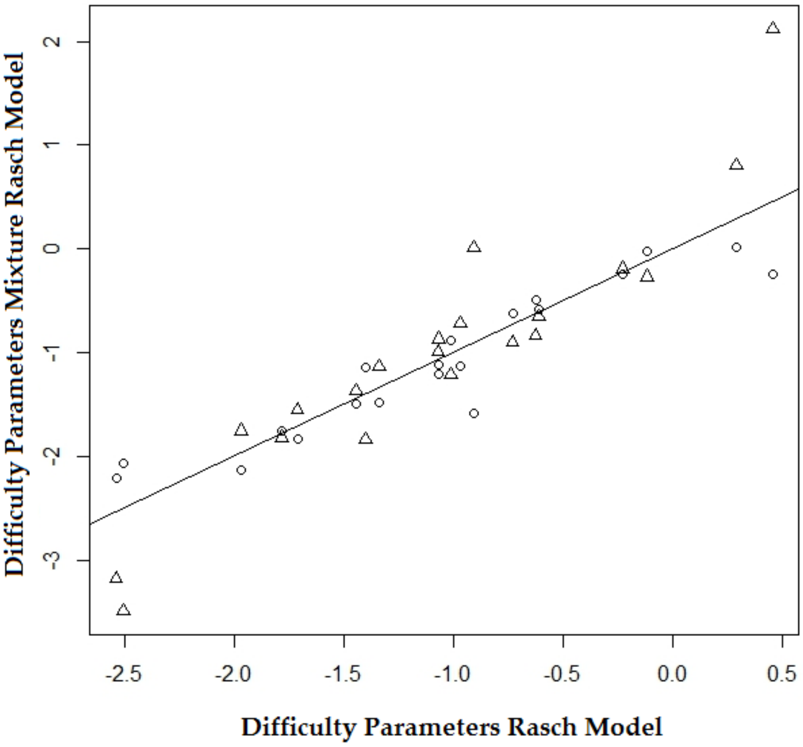


Figure 2.5 *Item difficulty parameters of the Rasch model and the mixture Rasch model for the latent class representing high-stakes responding (○) and the latent class representing low-stakes responding (Δ) estimated in the external anchor design.*

The difficulty parameters of both latent classes were compared between the mixture Rasch model and the Rasch model. The difficulty parameters of both models are comparable because the mean of the proficiency distribution of the 2009 operational test was fixed at 0 in both models. Figure 2.4 shows the estimated difficulty parameters of the internal anchor items for both the Rasch model and the mixture Rasch model. Interestingly, the difficulty parameters of the Rasch model and the mixture Rasch model were approximately the same for both latent classes, except for one item. It could be said that the latent classes were identified for the larger part by this particular item. Figure 2.5 shows the difficulty parameters of the external anchor items for both models. The differences between the difficulty parameters estimated using the Rasch model and the mixture Rasch model did not show a regular pattern. Because of the large number of items in the pre-test design, Figure 2.6 shows the difficulty parameters of the Rasch model and the mixture Rasch model separately for each latent class. The difficulty parameters estimated in the latent class representing high-stakes responding were approximately the same for the mixture Rasch model and the simple Rasch model, whereas the difficulty

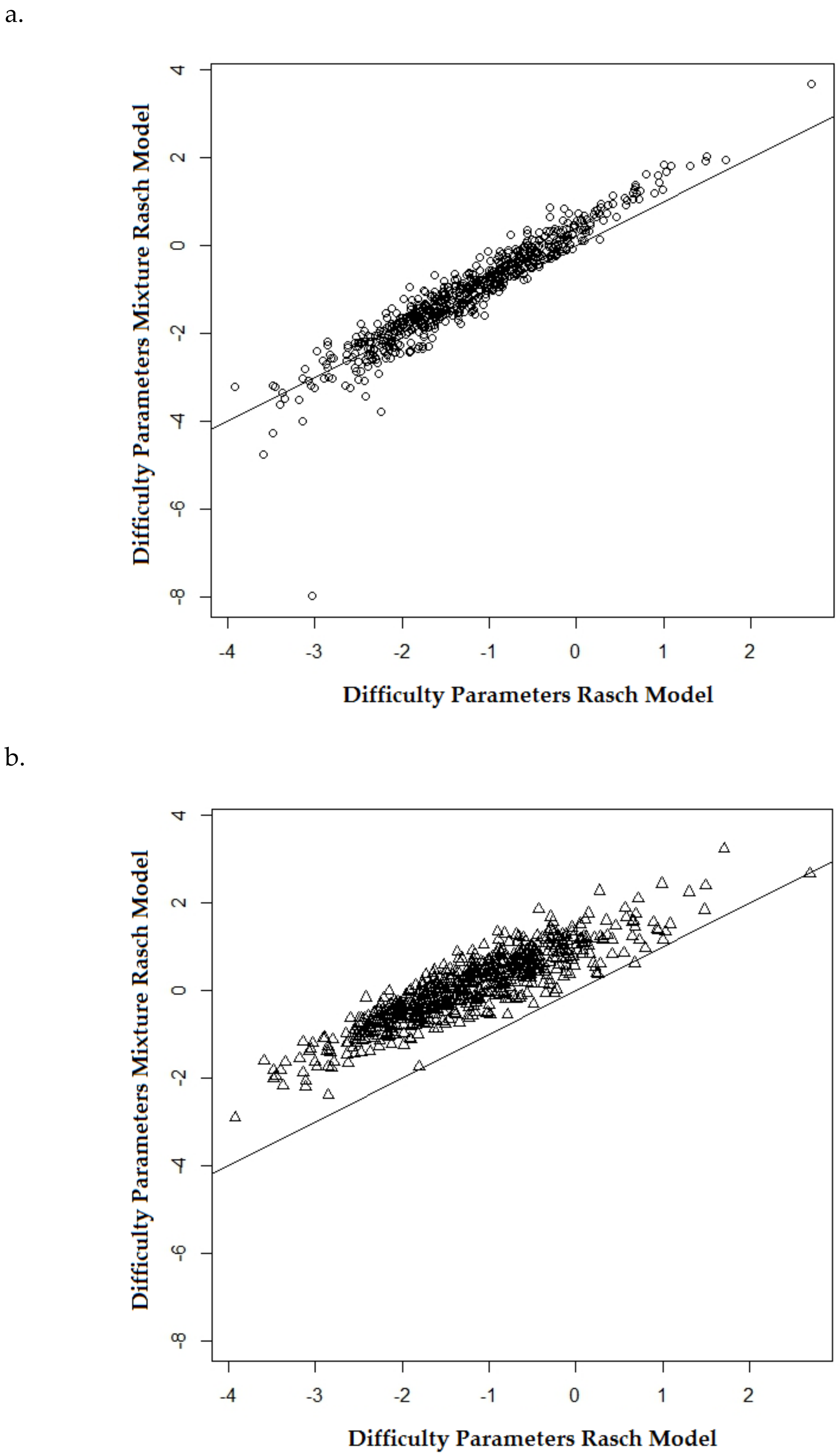


Figure 2.6 *Item difficulty parameters of the Rasch model and the mixture Rasch model for the latent class representing high-stakes responding (a) and the latent class representing low-stakes responding (b) estimated in the pre-test design.*

parameters estimated in the latent class representing low-stakes responding were higher for the mixture Rasch model and the simple Rasch model. As expected, for the pre-test design, the difficulty parameters were higher for the latent class representing low-stakes responding than the latent class representing high-stakes responding.

Differential Item Functioning

Five items were identified showing DIF between low-stakes and high-stakes administration conditions. After removal of these items, the Rasch model was fitted to the remaining data. The mean proficiencies of the 2009 and 2010 operational tests were 0.000 and 0.370,respectively, which is comparable to the linking result for the pre-test design using the Rasch model including the five DIF items. It was concluded that removal of items displaying DIF did not have an effect on the differences in estimated proficiency between the operational tests.

Discussion

We conclude that the result of the linking procedure depends on the type of data collection design. Specifically, the linking results differed between data collection designs, which differ with respect to the administration condition of the anchor items. Even though the external anchor design and the pre-test design both used anchor items administered in low-stakes administration conditions, the linking results for these two designs differed. Using the pre-test design to link the operational tests resulted in a substantial overestimation of the difference between the estimated proficiency of the populations administered the operational tests. Removing items showing DIF between high-stakes and low-stakes administration conditions did not improve the linking result for the pre-test design.

We also found evidence for the existence of differently motivated subpopulations. As a result of fitting the mixture Rasch model, for an external anchor, we found that the differences in estimated mean proficiency between the two operational tests did not change but that the differences had disappeared for the pre-test data. Class-membership probabilities of the latent class representing high-stakes responding might explain these results. The class-membership probabilities of the external anchor were almost the same in both years. Mittelhaëuser, Beguin, and Sijtsma (in press) argue that the linking result is only threatened if the effect of differential motivation is unequal between the populations. However, the class-membership probabilities of the pre-test booklets varied greatly, which rendered fitting a mixture Rasch model worthwhile.

The use of a mixture Rasch model proved to be useful with the current data. However, since the assumption of equal discrimination for all items is unlikely to be met in most real-data sets, it might be interesting to add varying discrimination parameters to the mixture model. Furthermore, the current research only investigated a mixture Rasch model with two latent classes, because it was assumed that examinees were either motivated or unmotivated to take the test. However, examinees could just as well have been motivated to a certain degree, in which case a model with more latent classes or a multidimensional IRT model is more appropriate to model item responding (Embretson & Reise, 2000).

Chapter 3

The Effect of Differential Motivation on IRT Linking Using the Rasch Model[*]

Abstract

The purpose of this study was to investigate whether simulated differential motivation between the stakes for operational tests and anchor items would produce an invalid linking result if the Rasch model was used to link the operational tests. This was done for an external anchor design and a variation of a pre-test design. The study also investigated whether a constrained mixture Rasch model could identify the latent classes in such a way that one latent class would represent high-stakes responding while the other would represent low-stakes responding. The results indicated that for an external anchor design, the Rasch linking result was only biased when the motivation level differed between the subpopulations that were presented with the anchor items. However, the mixture Rasch model could not identify the classes representing low-stakes and high-stakes responding. When a pre-test design was used to link the operational tests by means of a Rasch model, bias in the linking result was found in each condition. The amount of bias increased when the percentage of students showing low-stakes responding on the anchor items increased. The mixture Rasch model was only able to identify the classes representing low-stakes and high-stakes responding under a limited number of conditions.

[*] This chapter has been submitted for publication

In order to prevent item disclosure and maintain fairness, many testing programs use a new test form for every major administration (Holland & Rubin, 1982). Test forms may differ with respect to difficulty, and as a result, without further adaptation, scores on different test forms may not be comparable. A linking procedure is an adaptation that can be used to develop a common metric between test forms and adjust different scales for differences in levels of difficulty. A linking design frequently used in educational testing is the common-item non-equivalent groups design (see, e.g., Kolen & Brennan, 2004). In this design, different test forms are administered to different populations, and test forms can be linked by means of common items, also referred to as anchor items, to develop a common metric. The validity of the linking result depends, for example, on the question of whether the anchor items measure the same attribute as the test forms to be linked, henceforth referred to as the operational tests (Beguin & Hanson, 2001). Other factors that threaten to invalidate the linking of the operational tests and the anchor items concern order effects or the differences between the administration condition of the operational tests and the anchor items (see, e.g., Cook & Petersen, 1987; Klein & Jarjoura, 1985).

Differential student motivation (Holland & Wightman, 1982) regarding test taking needs to be considered in determining whether the anchor items measure the same attribute as the operational tests. Differential motivation refers to the difference in test-taking motivation between high-stakes and low-stakes (i.e., non-consequential) administration conditions. In a high-stakes administration condition, an examinee is expected to work harder and strive for maximum performance whereas a low-stakes administration condition elicits typical, rather than maximum, performance. In practice, testing programs might use anchor items administered in low-stakes conditions to link operational tests administered in high-stakes conditions (Wise & Kong, 2005). Unfortunately, if no important personal consequences are associated with test performance, students may care little whether their test scores accurately reflect their level on the attribute of interest (Reise & Flannery, 1996; Sundre, 1999; Wolf, Smith, & Birnbaum, 1995), and differences between student effort and, therefore, student performance in high-stakes and low-stakes administration conditions are expected (Wise & DeMars, 2005). The inconsistency in the stakes of the administration condition may lead to the misfit of the item response theory (IRT; see Van der Linden & Hambleton, 1997) model used for linking, or it may bias the linking result (Béguin, 2005).

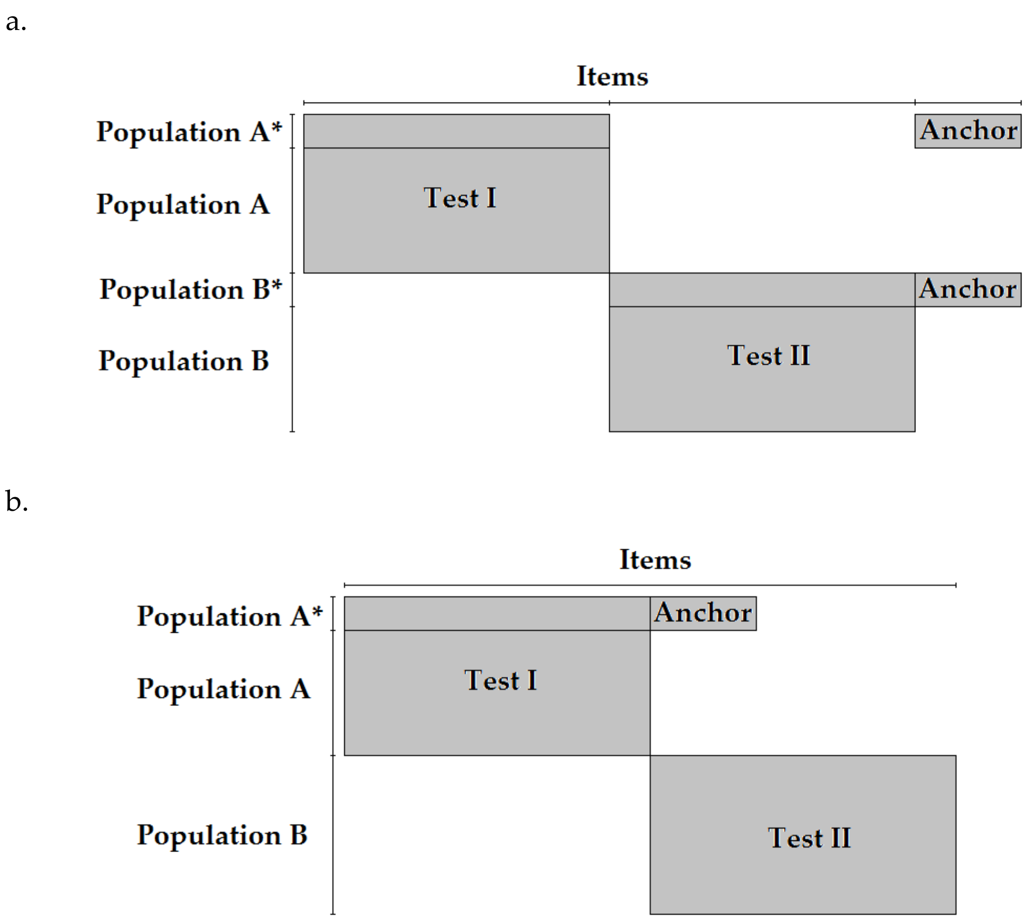


Figure 3.1 *(a) An external anchor design, (b) a pre-test design.*

We used a simulation study to investigate the extent to which differences between the stakes in the operational tests and the anchor items would produce an invalid linking result if the Rasch model was used to link two operational tests; we also investigated which circumstances affected the bias in the linking result.

Two types of data collection designs discussed here are the external anchor design and a variation of a pre-test design. Figures 3.1a and 3.1b show representations of an external anchor design and a pre-test design, respectively. The rows correspond to student data and the columns to item data. The boxes represent combinations of students and items for which data are available. In both data collection designs, population A is administered operational test I, and population B is administered operational test II. Since we focus on common-item non-equivalent groups designs, we do not assume equivalence of the populations, and anchor items are needed to link operational tests I and II.

In the external anchor design, the anchor items are administered to a subpopulation of population A (i.e., A*) and to a subpopulation of population B (i.e., B*). The anchor items are administered in addition to the two operational tests, and the total score on the anchor items does not contribute to the score on the operational tests; hence, the anchor items are referred to as external anchor items. Differences with respect to difficulty between the operational tests can be identified from the relative performance of subpopulations A* and B* on the anchor items. If the external anchor items are administered under a condition where the stakes are lower than the stakes of the operational tests, the validity of the linking result may be threatened. Theoretically, this is true only if the effect of differential motivation is unequal between the populations (Mittelhaëuser, Béguin, & Sijtsma, in press). To clarify this statement, one may first assume that the effect of differential motivation is equal among the populations administered with the external anchor items so that in comparison with the operational tests, the difficulty of the external anchor items is overestimated; but the relative performance of both populations on the external anchor items represents the true difference between population proficiency levels. Therefore, the linking result is unbiased. However, if the effect of differential motivation is unequal between the populations, the performance differences between the subpopulations on the anchor items due to differences in administration conditions will be attributed to differences between population proficiency levels, thereby resulting in an overestimation of the proficiency differences between the populations.

In a pre-test design, operational tests are linked using anchor items that are external to one of the operational tests (i.e., operational test I) and internal to the other operational test (i.e., operational test II). In this design, subsets of items intended for use in a new operational test (i.e., operational test II) are pre-tested on a subpopulation of students (i.e., A*) to whom the old operational test (i.e., operational test I) was administered. Differences with respect to the difficulty between the operational tests can be identified from the relative performance of subpopulation A* and population B on the anchor items. If the condition in which the anchor items are administered to subpopulation A* and population B differ in the sense that the stakes are lower for subpopulation A*, then subpopulation A* will likely show less effort and a correspondingly lower performance on the anchor items compared to population B. The performance differences between subpopulation A* and population B on the anchor items due to differences in administration conditions will be attributed to differences between population proficiency levels, thus resulting in an overestimation of the proficiency of population B.

In the current simulation study, the effect of differential motivation is operationalized by simulating the data on the anchor items in two latent classes, one representing high-stakes responding and the other representing low-stakes responding. In mixture IRT models, it is assumed that the data are a mixture of different datasets from two or more unidentified populations (Rost, 1997; Von Davier & Yamamoto, 2004), also called latent classes. Thus, theoretically, an IRT linking procedure using a mixture Rasch model is more robust against responding differences between different subgroups of students on the anchor items than a simple Rasch model.

If different latent classes produced the data, different model parameters would be valid for different subpopulations and a mixture IRT model could be used to identify the latent classes. For example, mixture IRT models have been used to identify subpopulations that differ with respect to the scalability of the items on personality traits (Rost, Carstensen, & Von Davier, 1997), to identify students who employ different solution strategies (Mislevy & Verhelst, 1987), and to identify known sources of contamination underlying item parameter estimates, such as test speededness (Bolt, Cohen, & Wollack, 2002; Yamamoto & Everson, 1997). Furthermore, a constrained mixture Rasch model can identify the latent classes in such a way that one latent class represents high-stakes response behavior while the other represents low-stakes response behavior (Béguin, 2005; Béguin & Maan, 2007).

The purpose of this study was to investigate whether simulated differences between the stakes of the administration conditions for the operational tests and the anchor items would produce an invalid linking result if the Rasch model was used to link the operational tests. We also investigated which circumstances affected the bias in the linking result. Since the data were generated from two latent classes that differed on the basis of response effort, the linking results obtained from a mixture Rasch model were used as a benchmark.

Method

Models

Let X_i denote the score on item i, and let k represent the number of items in a test. According to the mixture IRT model, the probability of passing item i,

$P(X_i = 1)$, depends on a class-specific person parameter θ_{jg}, which denotes the proficiency of student j when he/she belongs to latent class g, and a class-specific difficulty parameter β_{ig}. The mixture Rasch model defines the conditional response probability as

$$p(X_{ij} = 1|\theta_{jg}) = \frac{\exp(\theta_{jg} - \beta_{ig})}{1 + \exp(\theta_{jg} - \beta_{ig})}.$$

When aggregated across items, the probability of obtaining an item-score vector $\boldsymbol{x}_j = \{x_{1j}, x_{2j}, \dots, x_{kj}\}$ given proficiency θ_j, and the membership of class g equals

$$p(\boldsymbol{x}_j|g) = \prod_{i=1}^{k} \frac{\exp[x_{ij}(\theta_{jg} - \beta_{ig})]}{1 + \exp(\theta_{jg} - \beta_{ig})}.$$

Let π_g denote the proportion of the population belonging to class g ($g = 1, \dots, G$), which is also known as the class probability. The probability that student j belongs to class g equals

$$p(g|\boldsymbol{x}_j) = \frac{\pi_g p(\boldsymbol{x}_j|g)}{\sum_{g=1}^{G} \pi_g p(\boldsymbol{x}_j|g)}. \tag{3.1}$$

Von Davier and Yamamoto (2004) discussed the estimation procedure for the mixture IRT model by means of the EM algorithm. The Rasch model can be interpreted as a special case of the mixture Rasch model in which $G = 1$.

Data

Dichotomous item scores were generated under 40 conditions, which differed on the basis of the following factors:

Design. The same notation was used for samples and populations. The data were generated for an external anchor design and a pre-test design, as presented in Figure 3.1. Each design contained two operational tests (I and II), each consisting of 60 items. For each operational test, a sample of 5,000 item-score vectors was generated. In the external anchor design, 1,000 item-score vectors containing 0/1 scores for 20 anchor items were generated for both subpopulation A* and subpopulation B*. In the pre-test design, 1,000 item-score vectors containing 0/1 scores for 20 anchor items were generated for subpopulation A*. Item-score vectors for the operational tests were generated as if they had originated from a high-stakes administration condition in which we assumed that all students showed high-stakes responding. This was operationalized by choosing $\pi_1 = 1$ and $\pi_2 = 0$ in Equation 3.1 for all item-score vectors of both operational tests, thus fixing their class membership. The item-score vectors of the anchor items were generated as if they had originated from a low-stakes administration condition in which we assumed that students showed either high-stakes responding or low-stakes responding.

Item parameters of the anchor items. Empirical studies showed that students are more motivated in high-stakes administration conditions than in low-stakes administration conditions (Sundre, 1999; Wolf, Smith, & Birnbaum, 1995) and that motivated students perform better than unmotivated students (Wise & DeMars, 2005). Within an IRT framework, this results in items having higher item difficulty estimates when items are administered in low-stakes administration conditions than when they are administered in high-stakes administration conditions. The degree to which the item parameters used for generating data differed between classes of students showing differences in responding on the anchor items varied, resulting in small-effect conditions and large-effect conditions. In the small-effect conditions, the item parameters of the anchor items differed slightly between students who responded differently; under the large-effect conditions, the differences between the item parameters were larger. The item parameters for the operational tests and anchor items were selected as follows:

1. Real data obtained from the administration of 60 items in a high-stakes condition at the end of Dutch primary education was fitted to the Rasch model. The same items were also administered in a low-stakes pre-test condition. Hence, data were available from which both "high-stakes item parameters" and "low-stakes item parameters" were estimated for 60 items. These 60 high-stakes real-data item parameter estimates were used to generate artificial data for both operational tests in both designs.
2. To generate artificial data for the anchor items, the 60 items were ordered according to the difference between the item parameter estimated from high-stakes administration data and the item parameter estimated from low-stakes administration data.
3. The 20 items showing the greatest difference between the two item parameter estimates were used to generate data for the anchor items in the large-effect conditions. The 20 items showing the smallest difference between the two item parameter estimates were used to generate data for the anchor items in the small-effect condition.

Table 3.1 contains the item parameter estimates used to generate data for all conditions.

Table 3.1 *Real-data item difficulty parameters used to generate artificial data.*

Large effect		Small effect		Unique operational
High-stakes	Low-stakes	High-stakes	Low-stakes	High-stakes
-0.570	-0.116	0.483	0.523	0.542
-1.113	-0.637	-0.180	-0.120	-0.549
0.281	0.770	-0.024	0.048	0.672
0.248	0.744	1.031	1.112	0.423
0.683	1.199	0.187	0.341	-0.212
0.772	1.314	0.313	0.478	-0.722
1.020	1.562	-0.425	-0.238	-1.259
0.752	1.308	-0.310	-0.115	-0.632
0.267	0.833	-0.627	-0.405	-0.662
-0.363	0.245	1.557	1.787	0.450
-0.953	-0.336	0.018	0.260	0.617
0.000	0.628	-0.393	-0.133	0.379
-0.952	-0.249	0.474	0.736	1.258
0.266	0.980	0.283	0.598	0.806
0.055	0.841	0.273	0.630	0.060
-0.992	-0.206	0.248	0.607	-0.349
-0.903	-0.109	0.137	0.497	-0.774
-0.632	0.164	-0.535	-0.159	-0.621
-1.501	-0.693	0.823	1.230	0.219
0.847	1.661	1.087	1.517	-0.161

Proficiency of the populations. The data generated resulted in conditions with equivalent normal proficiency distributions for both populations ($\mu_A = 0$; $\mu_B = 0$; $\sigma_A = 1$; $\sigma_B = 1$) and conditions in which the means of the normal proficiency distributions differed by 0.25 standard deviations ($\mu_A = 0$; $\mu_B = 0.25$; $\sigma_A = 1$; $\sigma_B = 1$).

Proportion of students showing low-stakes responding on anchor items. Varying the proportion of students showing low-stakes responding on the anchor items resulted in conditions in which 10%, 25%, 50%, and 75% of the item-score vectors for the anchor items were generated using low-stakes item parameters. In the external anchor design, the proportion of students showing low-stakes responding on the anchor items varied across populations. This resulted in conditions whereby 50% of the students in subpopulation A* invested low-stakes responding on the anchor items whereas either 10% or 75% of the students in subpopulation B* invested low-stakes responding on the anchor items.

This simulation design resulted in 24 conditions for the external anchor design and 16 conditions for the pre-test design. For each of the 40 conditions, 1,000 datasets were generated, denoted by replication. In each replication h ($h = 1, \ldots, H$), operational tests I and II were linked by means of a Rasch model and a mixture Rasch model with ordinal constraints on the latent classes.

Analyses

Software. The OPLM software (Verhelst, Glas, & Verstralen, 1995) produced marginal maximum likelihood estimates of the item parameters from the Rasch model. An adapted version of OPLM used a mixture Rasch model to estimate the item parameters. The item-score vectors for the operational tests were modelled to exclusively belong to the first latent class, which was done by choosing $\pi_1 = 1$ and $\pi_2 = 0$ in Equation 3.1 for all item-score vectors for the operational tests. The item-score vectors for the anchor items could belong either to the first or the second latent class. Using these constraints on the latent classes facilitated the identification of latent classes such that one represented high-stakes responding and the other low-stakes responding (Béguin, 2005; Béguin & Maan, 2007).

For both the Rasch and mixture Rasch models, normally distributed proficiency distributions were approximated using Gauss-Hermite quadrature (Abramowitz & Stegun, 1972) with 180 quadrature points. Both designs were estimated using two marginal proficiency distributions. In the external anchor design, proficiency distributions were estimated for population A and subpopulation A* on the one hand and population B and subpopulation B* on the other hand. In the pre-test design, proficiency distributions were estimated for population A and subpopulation A* on the one hand and population B on the other hand.

Evaluation criteria

The criteria used to evaluate the performance of the estimation procedure and the linking precision in the 40 conditions were based on IRT observed-score equating. In IRT observed-score equating, an IRT model is used to estimate the distribution of observed number-correct scores (Kolen & Brennan, 2004; Zeng & Kolen, 1995), henceforth referred to as an estimated score distribution. In practice, the interest is on determining how population B would have performed on operational test I (or vice versa, how population A would have performed on operational test II). Two criteria were used to evaluate the equating precision in each condition. Given the item and population parameters used for generating data with the same estimated score distribution, and given the item and population parameters estimated in replication h, these criteria compared the estimated score distribution for population B on operational test I. Since the estimated score distributions were needed to compute the evaluation criteria used in this study, we shall first explain how the estimated score distributions were computed. Following this, we shall discuss the two evaluation criteria.

Estimating score distributions. The score distribution for students belonging to population A, with a normal proficiency distribution with a mean of μ_A, and a standard deviation of σ_A were computed by integrating over the population distribution of θ, that is,

$$f_r(x) = \int \sum_{\{x|r\}} f_r(x|\theta) g(\theta|\mu_A, \sigma_A) d\theta,$$

where $\{x|r\}$ in the summand stands for the set of all item-score vectors resulting in the same total score r. At each of the quadrature points, a recursion formula proposed by Lord and Wingersky (1984) was used to obtain $f_r(x|\theta)$, which is the score distribution of students with the same proficiency θ. A distinction was made between computing the estimated score distribution based on the item and population parameters used to generate the data (henceforth called true estimated score distribution) and computing the estimated score distribution based on the item and population parameters estimated in replication h.

Criterion 1. To compare the true estimated score distribution of population B on operational test I with the estimated score distributions resulting from the H replications, a mean squared error (MSE) was computed. Let $f_{true,r}$ be the frequency of score r based on the parameters used to generate the data. Let f_{hr} be the frequency of score r based on the parameters estimated in replication h. The mean MSE across scores was computed from the squared deviations of the two frequencies obtained in each combination of the H replications and the $k + 1$ scores, that is,

$$MSE = \frac{1}{H(k+1)} \sum_{h=1}^{H} \sum_{r=0}^{k} \left(f_{hr} - f_{true,r}\right)^2.$$

Furthermore, the MSE was decomposed into a term representing the mean across scores of the squared bias (mean bias) and a term representing the mean across scores of the variance (mean variance),

$$Mean\ bias = \frac{1}{k+1}\sum_{r=0}^{k}\left(\bar{f}_r - f_{true,r}\right)^2,$$

$$Mean\ variance = \frac{1}{H(k+1)}\sum_{h=1}^{H}\sum_{r=0}^{k}(f_{hr} - \bar{f}_r)^2\,,$$

where $\bar{f}_r$ is the mean frequency of score *r* across replications, that is,

$$\bar{f}_r = \frac{1}{H}\sum_{h=1}^{H} f_{hr}.$$

Criterion 2. The second criterion was based on comparing equivalent scores from the observed-score equating function with the true equivalent scores (i.e., equivalent scores based on the item and population parameters used to generate the data). Thus, equivalent scores based on the item and population parameters from replication h were compared with the equivalent scores based on the item and population parameters used to generate the data. Let $s_{true,r}$ be the integer score on operational test I, which is equivalent to score r on operational test II, based on equipercentile equating computed using the true item parameters and true latent proficiency distribution for population B. Let s_{hr} be the corresponding score estimated in replication h. Furthermore, let $p_{r,true}$ be the probability for population B to obtain score *r* on operational test I based on the true parameter values. To compare the equivalent scores, a weighted mean squared error (WMSE) was computed by summation across samples and scores. The scores were weighted by $p_{h,true}$, which resulted in

$$WMSE = \frac{1}{H}\sum_{r=0}^{k} p_{r,true}\sum_{h=1}^{H}(s_{hr} - s_{true,r})^2. \qquad (3.2)$$

The WMSE was decomposed into a term representing the weighted sum of the squared bias (weighted bias) of equated scores and a term representing the weighted sum of the variance (weighted variance) of the equated scores, therefore,

$$Weighted\ bias = \sum_{r=0}^{k} p_{r,true}\left(\bar{s}_r - s_{true,r}\right)^2,$$

$$Weighted\ variance = \frac{1}{H}\sum_{r=0}^{k} p_{r,true} \sum_{h=1}^{H} (s_{hr} - \bar{s_r})^2,$$

where $\bar{s_r}$ is the mean equivalent score of score r across replications, that is,

$$\bar{s_r} = \frac{1}{H}\sum_{h=1}^{H} s_{hr}.$$

The weighted mean absolute error (WMAE) is obtained if the squared error in Equation 3.2 is replaced by the absolute value of the error and, as a result,

$$WMAE = \frac{1}{H}\sum_{r=0}^{k} p_{r,true} \sum_{h=1}^{H} |s_{hr} - s_{true,r}|.$$

Compared to the WMSE, the WMAE can be interpreted more easily in terms of the deviation from $s_{true,r}$.

Results

The evaluation criteria for all conditions are presented in Table 3.2 through Table 3.5. In each table, the first column indicates whether the Rasch or mixture Rasch model was used to link the two tests. The second column (i.e., Pop) indicates whether data were generated from two populations with equivalent proficiency distributions ("0.00" mean difference) or whether data were generated using proficiency distributions with a 0.25 standard deviation mean difference. The third column gives the percentage of students generated to show low-stakes responding on the anchor items in subpopulations A* and B* (for the external anchor design) and subpopulation A* (for the pre-test design). Columns 4 and 5 give the MSE and the WMSE, respectively, with the amount of bias (as opposed to variance) in parentheses.

External anchor design

Small effect. Table 3.2 (small effect size) shows that the linking result was unbiased for the Rasch model if the effect of differential motivation was the same for subpopulations A* and B*. Since bias in the MSE and the WMSE was negligible, both could be attributed completely to variance across replications (not tabulated). Bias in the linking result increased as differences in motivation were larger. Differences were negligible between proficiency distributions that coincided and proficiency distributions having a 0.25 standard-deviation mean difference. The WMAE was approximately equal to the WMSE. When the linking result was biased, the differences between the equated scores based on the true item parameters and the equated scores based on the item parameters

Table 3.2 *Evaluation criteria for the external anchor design and small effect size.*

Model	Pop	% L-S responding A*-B*	Evaluation criteria MSE (bias)	WMSE (bias)	WMAE
Rasch	0.00	10-10	9.6 (0.0)	0.1 (0.0)	0.1
	0.25		9.7 (0.0)	0.1 (0.0)	0.1
	0.00	25-25	10.0 (0.0)	0.1 (0.0)	0.1
	0.25		9.7 (0.0)	0.1 (0.0)	0.1
	0.00	50-50	10.1 (0.0)	0.1 (0.0)	0.1
	0.25		10.2 (0.0)	0.1 (0.0)	0.1
	0.00	75-75	10.4 (0.0)	0.1 (0.0)	0.1
	0.25		10.1 (0.0)	0.1 (0.0)	0.1
	0.00	50-75	49.5 (38.2)	0.8 (0.6)	0.7
	0.25		51.1 (41.1)	0.8 (0.6)	0.8
	0.00	50-10	108.9 (98.3)	1.5 (1.3)	1.2
	0.25		111.7 (101.1)	1.5 (1.3)	1.1
Mixture Rasch	0.00	10-10	11.4 (0.0)	0.2 (0.0)	0.2
	0.25		11.4 (0.0)	0.2 (0.0)	0.2
	0.00	25-25	12.7 (0.0)	0.2 (0.0)	0.2
	0.25		12.6 (0.0)	0.2 (0.0)	0.2
	0.00	50-50	13.4 (0.0)	0.2 (0.0)	0.2
	0.25		12.6 (0.0)	0.2 (0.0)	0.2
	0.00	75-75	12.7 (0.0)	0.2 (0.0)	0.2
	0.25		13.2 (0.0)	0.2 (0.0)	0.2
	0.00	50-75	51.1 (36.9)	0.8 (0.5)	0.7
	0.25		50.5 (37.5)	0.8 (0.6)	0.7
	0.00	50-10	107.1 (93.4)	1.5 (1.3)	1.1
	0.25		110.6 (96.8)	1.5 (1.3)	1.1

Note: Pop = the difference between the mean of the proficiency distributions used to generate the data; % L-S responding A*-B* = the percentage of students generated to show low-stakes responding in subpopulation A* - subpopulation B*

from the replications were approximately 1 score unit.

Compared to the Rasch model, the mixture Rasch model did not produce substantially lower levels of bias in the linking result. The estimated class proportions for the latent class representing low-stakes responding were approximately .67 (varying across conditions from .66 to 68) for subpopulation A* and .67 (.66 to .68) for subpopulation B*. This result suggests that even though the effect size in the small-effect condition was large enough to produce

Table 3.3 *Evaluation criteria for the external anchor design and large effect size.*

Model	Pop	% L-S responding A*-B*	Evaluation criteria MSE (bias)	WMSE (bias)	WMAE
Rasch	0.00	10-10	10.8 (0.0)	0.1 (0.0)	0.1
	0.25		10.0 (0.0)	0.1 (0.0)	0.1
	0.00	25-25	12.1 (0.1)	0.2 (0.0)	0.2
	0.25		12.0 (0.1)	0.2 (0.0)	0.2
	0.00	50-50	12.5 (0.0)	0.2 (0.0)	0.2
	0.25		12.3 (0.1)	0.2 (0.0)	0.2
	0.00	75-75	12.1 (0.0)	0.2 (0.0)	0.2
	0.25		12.8 (0.2)	0.2 (0.0)	0.2
	0.00	50-75	290.7 (278.8)	4.1 (3.9)	2.0
	0.25		298.5 (286.9)	4.2 (4.0)	2.0
	0.00	50-10	709.7 (697.7)	10.0 (9.7)	3.1
	0.25		723.6 (711.9)	9.6 (9.4)	3.0
Mixture Rasch	0.00	10-10	17.7 (0.0)	0.3 (0.0)	0.3
	0.25		17.1 (0.0)	0.3 (0.0)	0.3
	0.00	25-25	26.0 (0.1)	0.4 (0.0)	0.4
	0.25		26.9 (0.0)	0.4 (0.0)	0.4
	0.00	50-50	42.8 (0.2)	0.7 (0.0)	0.6
	0.25		42.4 (0.1)	0.6 (0.0)	0.6
	0.00	75-75	42.6 (0.2)	0.7 (0.0)	0.5
	0.25		50.3 (0.3)	0.7 (0.0)	0.6
	0.00	50-75	195.8 (151.1)	2.8 (2.1)	1.5
	0.25		216.5 (166.0)	3.1 (2.3)	1.5
	0.00	50-10	494.2 (463.6)	7.0 (6.5)	2.5
	0.25		510.9 (480.1)	6.9 (6.4)	2.5

Note: Pop = the difference between the mean of the proficiency distributions used to generate the data; % L-S responding A*-B* = the percentage of students generated to show low-stakes responding in subpopulation A* - subpopulation B*

bias in the linking result in some conditions, it was too small for the estimation procedure in the mixture Rasch model to correctly identify the latent classes.

Large effect. Table 3.3 (large effect size) shows that comparable to the small effect size (Table 3.2), for the Rasch model, the linking result was minimally biased when differential motivation was equal for subpopulations A* and B*. Since bias in the MSE and the WMSE was small, both could be attributed largely to variance across replications (not tabulated). Bias in the linking result increased as motivation differences widened. Again, differences were negligible between conditions in which the proficiency distributions coincided and conditions in which the proficiency distributions differed with a 0.25 standard-deviation mean difference. When the linking results were biased, the differences between the equated scores based on the true item parameters and the equated scores based on the item parameters from the replications were approximately 2 and 3 score units in the condition where 75% and 10% of the students in subpopulation B* showed low-stakes responding, respectively.

For the conditions in which the Rasch model produced bias, the mixture Rasch model produced a somewhat smaller amount of bias. For conditions in which motivation differences between populations were generated, compared tothe Rasch model, the mixture Rasch model produced a score difference between the equated scores based on the true item parameters and the equated scores based on the item parameters from replications that were approximately 0.5 score units lower. The question remains whether the estimated latent classes represented high-stakes responding and low-stakes responding. Again, the estimated class proportions led us to believe that this was not the case. For example, when data were generated from equal proficiency distributions, and the percentage of low-stakes responding for subpopulations A* and B* was 50% and 10%, respectively, the estimated class proportions for subpopulation A* and B* were .40 and .31, respectively. This result suggests that despite the large effect size, the estimation procedure for the mixture Rasch model was still not able to correctly identify the latent classes.

Pre-test design

Small effect. Table 3.4 shows that for the Rasch model, bias increased as a percentage of low-stakes responding on the anchor items increased. For 50% and 75% low-stakes responding, the MSE and WMSE were largely due to bias. Substantial differences were absent between equivalent proficiency distributions and proficiency distributions with a 0.25 standard-deviation mean difference. The difference between the WMAE and the WMSE increased as the percentage of low-stakes responding increased. The differences between the equated scores based on the true item parameters and the equated scores based on the item parameters from the replications increased from 0.1 to almost 2 score units as the percentage of low-stakes responding increased.

Compared to the Rasch model, the mixture Rasch model produced a similar or smaller amount of bias. For the latent class representing low-stakes

Table 3.4 *Evaluation criteria for the pre-test design and small effect size.*

Model	Pop	% L-S responding A*	Evaluation criteria MSE (bias)	WMSE (bias)	WMAE
Rasch	0.00	10	14.5 (6.8)	0.1 (0.0)	0.1
	0.25		12.7 (5.2)	0.1 (0.1)	0.1
	0.00	25	46.9 (39.5)	0.5 (0.3)	0.5
	0.25		46.9 (39.7)	0.5 (0.3)	0.5
	0.00	50	161.6 (154.2)	1.5 (1.4)	1.2
	0.25		163.7 (156.0)	1.5 (1.4)	1.2
	0.00	75	357.1 (348.9)	3.9 (3.8)	1.9
	0.25		363.7 (355.7)	3.9 (3.8)	1.9
Mixture Rasch	0.00	10	21.2 (1.5)	0.3 (0.0)	0.3
	0.25		18.5 (0.8)	0.3 (0.1)	0.3
	0.00	25	41.2 (17.8)	0.5 (0.1)	0.4
	0.25		44.8 (18.2)	0.5 (0.1)	0.5
	0.00	50	116.5 (80.6)	1.2 (0.7)	0.9
	0.25		114.3 (80.9)	1.2 (0.7)	0.9
	0.00	75	267.8 (234.6)	3.0 (2.5)	1.6
	0.25		272.7 (240.9)	3.0(2.5)	1.6

Note: Pop = the difference between the mean of the proficiency distributions used to generate the data; % L-S responding A* = the percentage of students generated to show low-stakes responding in subpopulation A*

responding, the estimated class proportions were approximately .11, .13, .18,and .26 for conditions where 10%, 25%, 50%, and 75% of the students showed low-stakes responding on the anchor items, respectively. The estimated class proportions were the same for conditions in which equal proficiency distributions and different proficiency distributions were used to generate data. As with the external anchor design, this led us to believe that even though the effect size in the small effect condition was large enough to produce substantial bias in the linking result in some conditions, the effect size was too small for the mixture Rasch model to correctly identify the latent classes in each condition.

Large effect. Table 3.5 shows that the Rasch model produced a higher level of bias as the percentage of students showing low-stakes responding on the anchor items increased. The MSE and WMSE were largely affected by bias. The results were approximately the same for equivalent proficiency distributions and different proficiency distributions. The difference between

Table 3.5 *Evaluation criteria for the pre-test design and large effect size.*

		% L-S responding	Evaluation criteria		
Model	Pop	A*	MSE (bias)	WMSE (bias)	WMAE
Rasch	0.00	10	51.9 (43.8)	1.0 (1.0)	1.0
	0.25		51.8 (42.8)	1.1 (1.0)	1.0
	0.00	25	288.1 (279.4)	4.8 (4.7)	2.1
	0.25		291.8 (283.1)	4.7 (4.5)	2.1
	0.00	50	1099.8 (1091.3)	17.0 (16.8)	4.1
	0.25		1136.5 (1126.7)	16.8 (16.7)	4.0
	0.00	75	2468.6 (2459.5)	37.4 (37.2)	6.1
	0.25		2530.0 (2521.1)	36.2 (36.0)	5.9
Mixture Rasch	0.00	10	43.1 (0.5)	0.7 (0.1)	0.6
	0.25		41.1 (1.4)	0.7 (0.2)	0.6
	0.00	25	72.5 (2.2)	1.2 (0.2)	0.8
	0.25		71.6 (1.4)	1.1 (0.1)	0.8
	0.00	50	139.5 (8.6)	2.2 (0.3)	1.2
	0.25		144.1 (11.7)	2.2 (0.4)	1.2
	0.00	75	679.9 (347.3)	10.6 (5.8)	2.6
	0.25		707.1 (374.5)	10.5 (6.0)	2.6

Note: Pop = the difference between the mean of the proficiency distributions used to generate the data; % L-S responding A* = the percentage of students generated to show low-stakes responding in subpopulation A*

the WMAE and the WMSE increased as the percentage of low-stakes responding on the anchor items increased. The differences between the equated scores based on the true item parameters and the equated scores based on the item parameters from the replications increased from one to approximately six score units as the percentage of low-stakes responding increased.

Compared with the Rasch model, the mixture Rasch model produced a substantially smaller amount of bias. The estimated class proportions for low-stakes responding were approximately .13, .23, .44, and .52 for conditions where 10%, 25%, 50%, and 75% of the students showed low-stakes responding, respectively. It seemed that for the first three conditions, the estimated class proportions approximated the actual class proportions used in generating data.

Discussion

We used simulated data to investigate the amount of bias introduced in the linking result when differential motivation affects the scores on the anchor items. As expected, for an external anchor design, the linking result was only biased when motivation differed between the subpopulations presented with the anchor items. The conclusions were the same for conditions in which equivalent proficiency distributions or proficiency distributions having different means were used to generate data. Since two latent classes were used to generate the data, the linking results from the mixture Rasch model served as a benchmark. However, the mixture Rasch model could only identify the classes representing low-stakes and high-stakes responding when a pre-test design was used to link the two tests, when the effect size was large, and when the percentage of students showing low-stakes responding was less than 75%. It is somewhat worrisome that under some conditions, the effect size was large enough to create substantial bias in the linking result but that the effect size was not large enough to control for this bias by means of a mixture IRT model with constraints on the latent classes.

Simulation studies only cover a limited variation of conditions, and the model used to generate the data does not perfectly represent real datasets (Davey, Nering, & Thompson, 1997). Therefore, the results should be interpreted with caution. For example, we operationalized differences in response efforts between high-stakes and low-stakes administration conditions by assigning students to one or two latent classes, respectively, in which the latent classes had different item-response probabilities. This operationalization of differential motivation may not realistically reflect differential motivation that is effective in real test taking.

The focus of the current study was on the Rasch model. Less restrictive IRT models than the Rasch model are available (Embretson & Reise, 2000). Such models are expected to better fit empirical data. However, the Rasch model has benefits, such as the sum score being a sufficient statistic for the proficiency estimate. Furthermore, research indicates that the Rasch model is robust under a variety of circumstances (Dinero & Haertel, 1977; Forsyth, Saisangjan, & Gilmer, 1981), thus suggesting that the use of more complex IRT models may not provide less biased results.

Chapter 4

Modeling Differences in Test-Taking Motivation: Exploring the Usefulness of the Mixture Rasch Model and Person-Fit Statistics[*]

Abstract

In analyzing test data, it is often assumed that students were motivated to answer the items correctly, hence that the attribute of interest drove test performance. However, if the test is administered in a low-stakes administration condition or if students do not receive feedback, students might not put their best effort into answering the items correctly. Within the item response theory (IRT) framework, lack of motivation threatens the correctness of proficiency and item parameter estimation and therefore the usefulness of the IRT model. The goal of the current study was to explore to what extent a mixture Rasch model and the l_z person-fit statistic could be used to model motivational differences in data administered in a low-stakes administration condition. In modeling the mixture Rasch model, constraints distinguished two latent classes of students: (1) a class representing "motivated" response behavior and (2) a class representing "unmotivated" response behavior. We investigated the usefulness of the mixture modeling strategy in a sample of primary-school students (N = 1,512) by comparing the posterior probabilities of the mixture Rasch model and the student's self-reported motivation. Furthermore, the study investigated the relationship between the student's self-reported motivation and the l_z person-fit statistic.

[*] This chapter has been published as: Mittelhaëuser, M., Béguin, A. A., & Sijtsma, K. (2013). Modeling differences in test-taking motivation: Exploring the usefulness of the mixture Rasch model and person-fit statistics. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.). *New developments in quantitative psychology* (pp. 357-370). New York: Springer.

Item response theory (IRT) models are useful in educational measurement for supporting the construction of measurement instruments, linking and equating of measurements, and evaluation of test bias (Scheerens, Glas, & Thomas, 2003). However, the IRT model must fit the data so as to be applicable to practical testing problems and yield consistent proficiency level and item parameter estimates. Unfortunately, researchers often implicitly assume that scores on a test are valid indicators of a student's best effort (Wolf & Smith, 1995) but Wainer (1993, p. 12) noted that: "If a test doesn't count for specific individuals, how can we be sure that they are trying as hard as they might if it mattered?". Over the years, evidence has accumulated that if item performance does not contribute to the test score or if no feedback is provided, students may not give their best effort and perform to their best ability (e.g., Wise & DeMars, 2005; O'Neill, Sugrue, & Baker, 1996; Kiplinger & Linn, 1996). Under-performance is typical for tests administered in a low-stakes administration condition. Consequently, performance on items administered in a low-stakes condition may differ from performance on items administered in a high-stakes condition, resulting in unusual patterns of item scores or in relatively poor performance on the low-stakes items. Within an IRT framework, low-stakes performance threatens the correct estimation of the proficiency and item parameters. For example, Mittelhaëuser, Béguin, and Sijtsma (2011) found that using low-stakes common items to link two high-stakes tests yielded different conclusions about the ability distributions compared to using high-stakes common items.

This article explores the usefulness of two methods that may be helpful in removing bias in parameter estimation caused by the low-stakes administration condition of a test. The first method uses a mixture Rasch model that assumes that the data are a mixture of different data sets from two or more latent populations (Rost, 1997; Von Davier & Yamamoto, 2004), also called latent classes. If the mixture assumption is correct, a Rasch model does not hold for the entire population but different model parameters are valid for different subpopulations. Let X_i denote the score on item i, and let k denote the number of items in the test. According to the mixture Rasch model, the probability of passing item i ($X_i = 1$) depends on a class-specific person parameter, θ_{jg}, which denotes the proficiency of student j if he/she belongs to latent class g. The conditional response probability is defined as:

$$P\left(X_i = 1 \middle| \theta_{jg}\right) = \frac{\exp(\theta_{jg} - \beta_{ig})}{1 + \exp(\theta_{jg} - \beta_{ig})}$$

where β_{ig} is a class-specific difficulty parameter. The probability of obtaining an item-score vector, $\boldsymbol{x} = \{x_1, x_2, \dots, x_k\}$, given proficiency θ_{jg} equals

$$P(\boldsymbol{x}_j|\theta_{jg}) = \prod_{i=1}^{k} \frac{\exp[x_i(\theta_{jg} - \beta_{ig})]}{1 + \exp(\theta_{jg} - \beta_{ig})}.$$

Let π_g denote the proportion of the population that belongs to class g ($g = 1, \dots, G$). The probability for student j to belong to class g, also known as the posterior probability, depends on the item-score vector; that is,

$$P(g|\boldsymbol{x}_j) = \frac{\pi_g P(\boldsymbol{x}_j|g)}{\sum_{g=1}^{G} \pi_g P(\boldsymbol{x}_j|g)}. \tag{4.1}$$

Mixture IRT models can be used to identify classes resulting from different types of response behavior. Consequently, the mixture strategy can also be used to handle known sources of contamination in item parameter estimates. For example, Bolt, Cohen, and Wollack (2002) used a mixture Rasch model with ordinal constraints to help remove the effect of test speededness on item parameter estimates. We used other constraints facilitating identification of latent classes such that one of the latent classes represents “high-stakes response behavior” and the other latent class “low-stakes response behavior” (Béguin, 2005; Béguin & Maan, 2007). As the probability for each individual to belong to a latent class (i.e., posterior probability) can be estimated, it is possible to identify the item-score vectors the low-stakes administration condition of the test affect. The posterior probabilities represent the probabilities for a student to respond in either a “low-stakes manner” or a “high-stakes manner.”

Alternatively, person-fit methods assign a value to each individual vector of item scores, and a statistical test is used to determine whether the underlying IRT model fits the item scores (Embretson & Reise, 2000). Significant person-fit values identify item-score vectors for which the IRT model does not fit, and the researcher may decide to remove the aberrant item-score vectors from the data set (Meijer & Sijtsma, 1995). The remaining set of item-score vectors for which the IRT model fits are expected to produce consistent parameter estimates. The l_z statistic is a well-known person-fit statistic (Drasgow, Levine, & Williams, 1985). By estimating the l_z statistic on a low-stakes item-score vector given the ability parameter estimated on a high-stakes test, it may be possible to detect students driven by unmotivated response behavior.

The goal of this study was to explore whether indicators of non-typical response behavior, such as the posterior probabilities from a mixture IRT model and the l_z person-fit statistic can be used to model motivational differences between students. We investigated the relationship between student's self-reported motivation on the one hand and the posterior probabilities of the mixture Rasch model and the l_z statistic on the other hand.

Method

Participants and Design

Four different scales were used to collect data: the End of Primary School Test 2012 (Eindtoets Basisonderwijs), the pre-test of the End of Primary School Test 2013, a scale measuring test-taking motivation, and a scale measuring social desirability. The order in which the different scales are discussed below corresponds to the order in which they were administered to the students.

Pre-test. Subsets of items intended for use in a high-stakes test are usually pre-tested on different samples of students to examine the statistical characteristics of the items before including them in a high-stakes test. To pre-test math items for the End of Primary School Test 2013, eighth-grade primary-school students (*N* = 9,124) were presented with a pre-test containing math items. Items most suitable for the population were selected for the End of Primary School Test 2013. Twenty-seven different pre-test versions also called test booklets were constructed, varying in test length from 30 to 60 items and including 585 multiple-choice items in total. The responses were coded 0 representing a wrong answer and 1 representing a right answer. The number of respondents per test booklet ranged from seven to 516. However, as a given pre-test item was administered in more than one pre-test booklet, the number of observations per item ranged from 332 to 1,424. The pre-test was used in most schools to practice for the high-stakes End of Primary School Test 2012, but the students were aware that they would not receive a score on the pre-test. Therefore, the pre-test is considered to be administered in a low-stakes administration condition.

Test-taking motivation. After the administration of the pre-test, a subsample of 1,512 students was administered a questionnaire containing nine items that measured test-taking motivation (TTM). The construction of the items was inspired by existing scales, such as the Test-Taking Motivation Questionnaire (Eklöf, 2006), the Student Opinion Scale (Thelk, Sundre, Horst, & Finney, 2009), and a subset of items from the self-report questionnaires of

Table 4.1*Test-taking motivation items with mean scores and component loadings*

Item		M	Loadings		
			A1	A2	A3
1	I enjoy going to school.	3.00	-.121	.061	**.853**
2	I enjoy learning math.	2.86	.011	-.107	**.817**
3	I did my best on the math items.	3.80	**.705**	.066	.103
4	My teacher wants me to do my best on the math items.	3.80	-.034	**.841**	.001
5	My parents want me to do my best on the math items.	3.83	.009	**.823**	.055
6	I did a good job on the math items.	3.21	**.525**	-.178	.205
7	The kids in my class did their best on the math items.	3.53	**.409**	.202	-.061
8	I could have worked harder on the math items.	2.88	**.788**	-.097	-.115
9	I'm curious about how many math items I answered correctly.	3.66	.203	.132	**.422**

the Education Quality Accountability Office (Zerpa, Hachey, van Barneveld, & Simon, 2011). Each item was answered on a 4-point Likert-scale (1 = No, 2 = Not so much, 3 = Kind of, 4 = Yes). Table 4.1 shows English translations of the items.

Social desirability. To check whether the tendency to answer in a socially desirable way influenced self-reported motivation, the students were administered six items stating desirable but uncommon behavior. The construction of the items was inspired by the Children's Social Desirability Scale (Baxter et al., 2004). Each item was answered as Not True (1) or True (2). Table 4.2 provides English translations of the items.

End of Primary School Test 2012. Each year in February, the End of Primary School Test is administered to students who are in the last year of Dutch primary education. The test results provide an independent advice to primary-school teachers, parents and secondary-schools about the most appropriate type of secondary education for a student. The test is administered in a high-stakes condition, and secrecy of the items is vital; hence, the test form is renewed each year. The End of Primary School Test 2012 contained 60 multiple-choice math items. The responses were coded 0 representing a wrong answer and 1 representing a right answer. In total, 144,708 students completed the math items of the End of Primary School Test 2012.

Table 4.2 *Social desirability items with mean scores and standard deviations*

Item		*M*	*SD*
1	I like all the kids in my class.	1.49	.50
2	I always tell the truth.	1.35	.48
3	I never fight.	1.15	.36
4	I always do what my teacher tells me to do.	1.60	.49
5	I always behave well.	1.44	.50
6	I never lie.	1.27	.45

Analyses

All analyses were performed using SPSS version 20 unless stated otherwise.

Principal components analysis. A principal components analysis (PCA) was performed to investigate the internal structure of the TTM scale. After motivational components of the TTM scale were identified, the reliability estimate known as the greatest lower bound (GLB) was calculated for the total TTM scale using Factor 8.1 (Lorenzo-Seva & Ferrando, 2006).

Mixture Rasch model. The data of the pre-test and the data of the End of Primary School Test 2012 were combined, providing 9,124 item-score vectors containing items administered in a low-stakes condition (pre-test items) and different items administered in a high-stakes condition (items from the End of Primary School Test 2012). A mixture Rasch model was estimated for this dataset using a dedicated version of the OPLM software (Verhelst, Glas, & Verstralen, 1995; Béguin, 2008). We anticipated that we would not find motivational differences in the high-stakes administration condition. Therefore, the item-score vectors of the End of Primary School Test 2012 were modeled as being exclusively part of the first latent class. This was done by setting $\pi_{g=0} = 0$ and $\pi_{g=1} = 1$ in Equation 4.1 for all item-score vectors of the End of Primary School Test 2012. The item-score vectors of the pre-test could be in either the first or the second latent class. To identify the model, it was assumed that student's abilities did not differ across latent classes.

After estimating the mixture Rasch model for the 9,124 item-score vectors, the posterior probabilities of the 1,512 students who completed the TTM scale were related to their self-reported motivation. This was done by estimating the correlation coefficient and inspecting the mean posterior probability for each separate TTM sum score. Furthermore, the item difficulty parameters of both latent classes were plotted to inspect the differences between the latent classes.

Person-fit. The l_z statistic (Drasgow et al, 1985; Meijer & Sijtsma, 1995) is a person-fit statistic that assesses the likelihood of an item-score vector under a specific IRT model. The l_z statistic is given by

$$l_z = \frac{l - E(l)}{\sqrt{Var(l)}}$$

where l denotes the unstandardized likelihood of the item-score vector and $E(l)$ and $Var(l)$ denote the expected likelihood and the variance of the likelihood, respectively. These three quantities are given by:

$$l = \sum_{i=1}^{k} \{X_i \ln P_i(\theta) + (1 - X_i) \ln[1 - P_i(\,\theta)]\}$$

with

$$E(l) = \sum_{i=1}^{k} \{P_i(\theta) \ln[P_i(\theta)] + [1 - P_i(\theta)] \ln[1 - P_i(\theta)]\}$$

and

$$Var(l) = \sum_{i=1}^{k} P_i(\theta)[1 - P_i(\theta)][\ln \frac{P_i(\theta)}{1 - P_i(\theta)}]^2.$$

The l_z statistic is assumed to be a standard normal deviate, with large negative values providing evidence of misfit.

For each student, the l_z statistic was calculated using the statistical program R (R Development Core Team, 2010) by means of the following three steps:

1. Item parameters of the Rasch model were estimated for the End of Primary School Test 2012 and pre-test concurrently.
2. The item parameters estimated in step 1 were fixed and the proficiency parameters of the Rasch model were estimated for the End of Primary School Test 2012.
3. The l_z statistic was calculated for the pre-test items, given the item parameters and proficiency parameters estimated in step 1 and 2, respectively.

The l_z statistic provided a likelihood measure of the low-stakes item-score pattern of the pre-test given the ability estimate based on the high-stakes item-score pattern of the End of Primary School Test 2012. High negative l_z values suggested motivational differences between the administration conditions.

After having estimated the l_z statistic for the 9,124 item-score vectors, the l_z statistics of the 1,512 students who completed the TTM scale were related to their self-reported motivation. This was done by estimating the correlation coefficient and analyzing the mean l_z statistic for each sum score on the TTM scale.

Social desirability. We used the Kruskal–Wallis Test to investigate the relationship between the score on the TTM scale and the social desirability (SD) scale so as to determine whether social desirability influenced the TTM scores.

Results

Principle components analysis

A PCA was performed on the nine items from the TTM scale. After 53 cases having missing values were deleted, the analysis was performed using data from 1,459 students. Prior to performing the PCA, the suitability of the data for PCA was assessed. Bartlett's Test of Sphericity (Bartlett, 1954) reached statistical significance and the Kaiser-Meyer-Oklin (Kaiser, 1974) value was .639, indicating that the data were suitable for PCA.

The PCA produced three components having eigenvalues exceeding 1 that explained 24.1%, 16.4%, and 12.7% of the variance, respectively. Analysis of the screeplot did not show a clear elbow. However, the loadings of the three-component solution revealed a simple structure. To aid the interpretation of the components, oblimin rotation was performed. The loadings are presented in Table 4.1, where the highest loadings per item are presented in boldface. The three-component solution explained a total of 53.2% of the variance. The first component can be interpreted as a "general TTM" component, the second component as an "external motivation" component, and the third component as measuring "general attitudes". The small number of items in each subscale renders the usefulness of the separate subscales that might be constructed based on these components limited. Therefore, we decided to use the sum score on the total TTM scale in all subsequent analyses. The GLB was calculated for the total TTM scale and equaled .71. This value suggests a reliability that allows less important decisions about individuals (Evers, Lucassen, Meijer & Sijtsma, 2010).

Mixture Rasch model

We computed the correlations between students' self-reported motivation and their posterior probabilities in a subsample of 1,453 students without incomplete data patterns. A significant but small positive relation between the variables was found, $r = .09$, $p = .001$.

We inspected the mean posterior probability for each sum score on the TTM scale. Figure 4.1a shows the results. Each of the sum scores of 16, 18, and 20 was produced by just one examinee, so that 95% confidence intervals for the mean posterior probabilities could not be determined. The student having the lowest score of 16 on the TTM scale had a very low posterior probability of belonging to the "motivated" class. However, the student having a sum score of 20 on the TTM scale had a very high posterior probability of belonging to the "motivated" class. The student having a TTM sum score of 16 indeed performed better on the high-stakes End of Primary School Test (95% of the items correctly answered) than on the low-stakes pre-test (41.67% of the items correctly answered). The student having a sum score of 20 performed better on the low-stakes pre-test (66.67% of the items correctly answered) than on the high-stakes End of Primary School Test (56.67% of the items correctly answered). The mean percentage of correctly answered items in the sample of 1,512 students on the End of Primary School Test was 72.56, and the mean percentage of correctly answered items on the pre-test was 64.96. It appears that the administration condition indeed influenced the student having a sum score of 16. Furthermore, the student having a sum score of 20 showed an average performance on the pre-test but scored below average on the End of Primary School Test.

As the number of observations on the lower sum scores on the TTM scale were very low (Sum score 16: n = 1, 18: n = 1, 20: n = 1, 21: n = 7, 22: n = 5, 23: n = 12), we combined the observations for the low sum scores. Figure 4.1b shows the relationship between the mean posterior probability and the TTM sum score.

The mean posterior probability was low for the lower TTM sum scores and stabilized starting from sum score 27 onward at a mean posterior probability of .6. The 95% confidence interval for the mean posterior probability for sum score 36 was slightly wider than the confidence intervals for sum score 27 onward.

Figure 4.2 shows the item difficulty parameters of both latent classes. The difficulty parameters for most items were higher in the "unmotivated" class than in the "motivated" class, which was expected. However, for a few items the difficulty parameters were higher in the "motivated" class than in the "unmotivated" class.

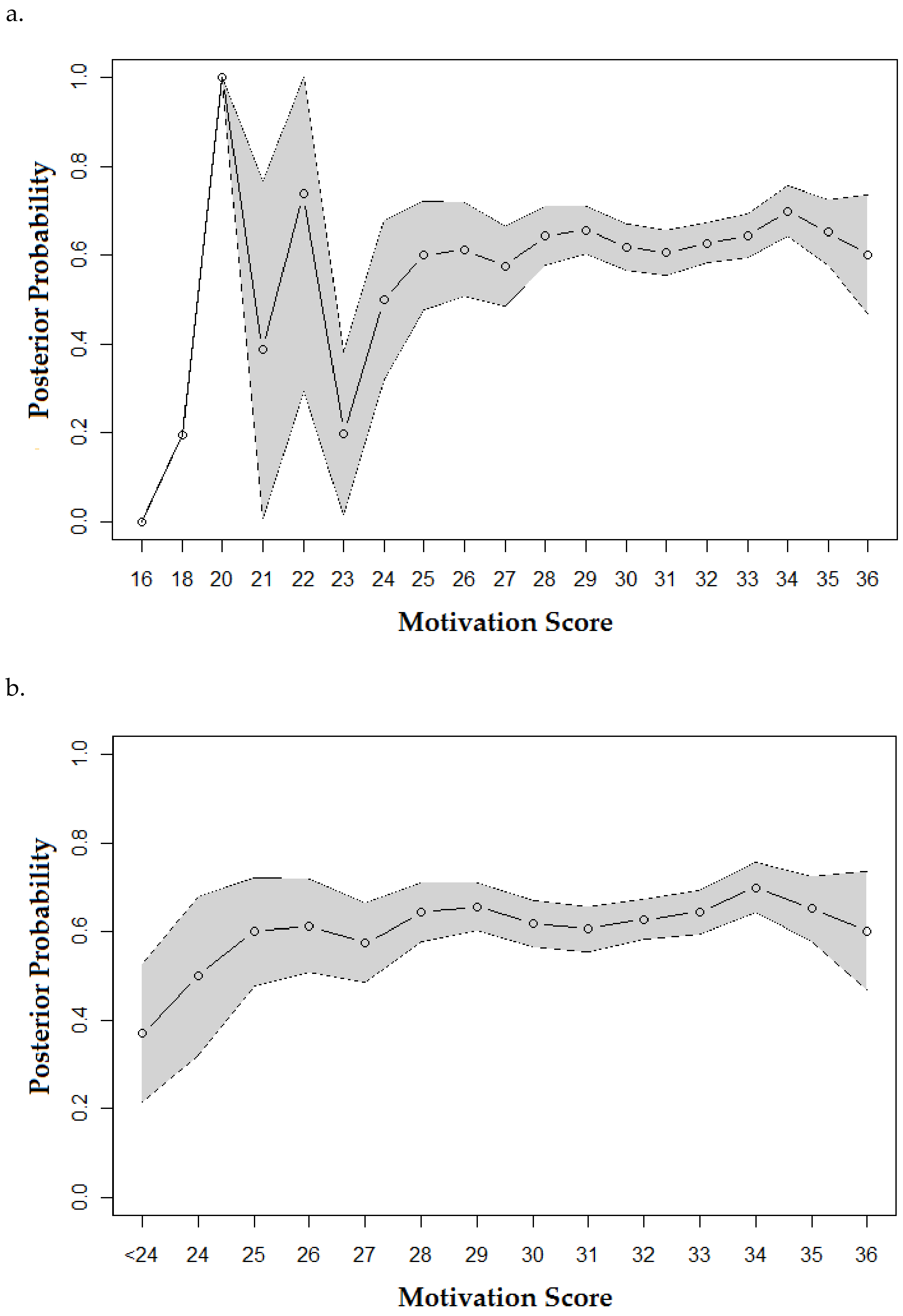


Figure 4.1 *(a) Mean posterior probability per motivation score, (b) mean posterior probability per motivation score with the lowest sum scores on the TTM scale combined. The gray area represents the 95% confidence interval for the mean posterior probability.*

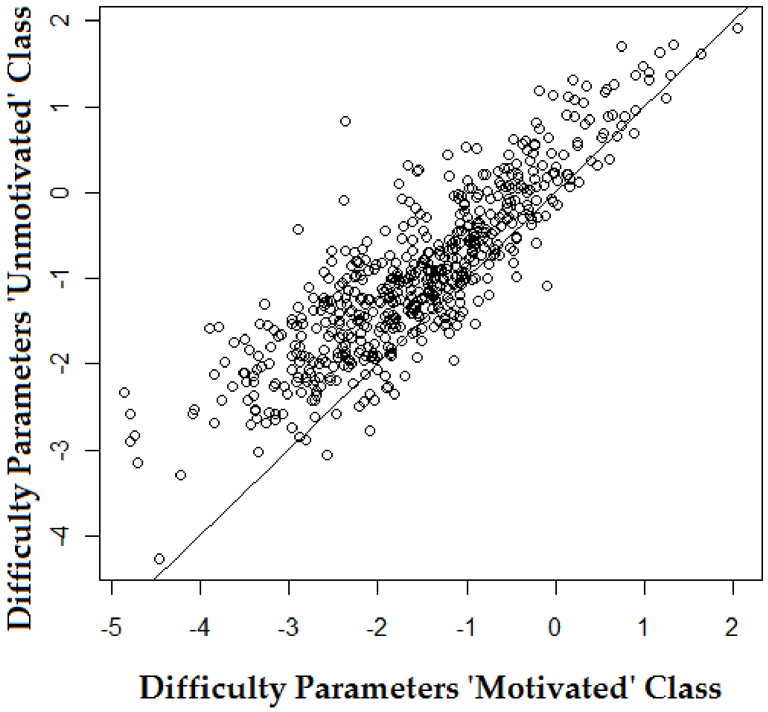


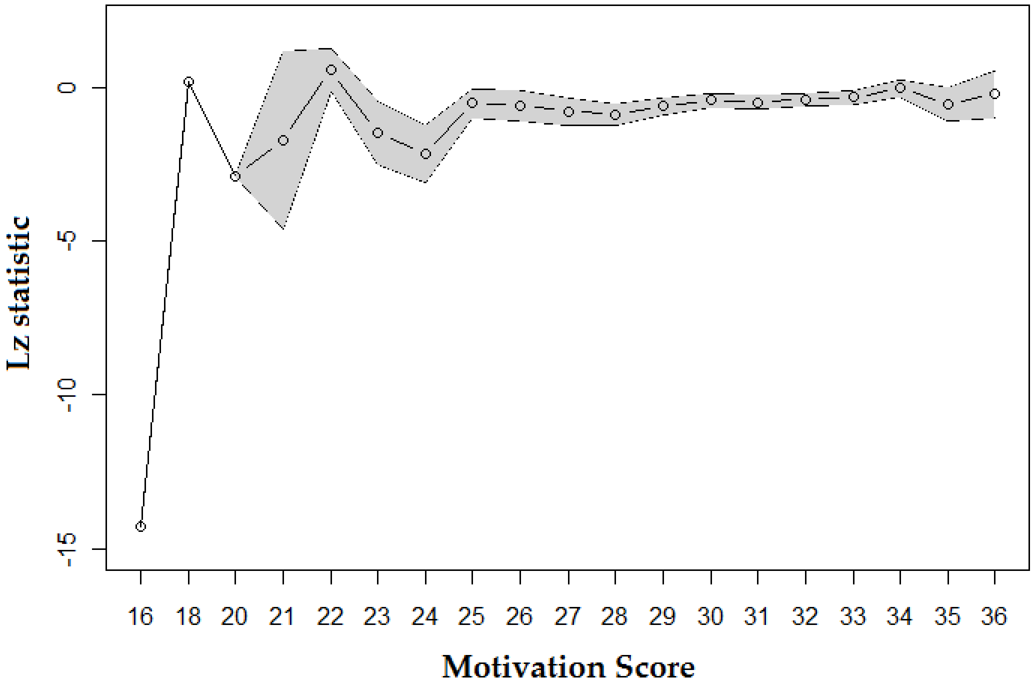
Figure 4.2 *Comparison of the item parameters estimated for the two latent classes*

Person-fit

The correlation between students' self-reported motivation and their l_z statistics equaled r = .15, p =< 0.001 (N = 1,453, incomplete cases removed). Figure 4.3a shows the mean l_z statistic for each TTM sum score. Due to the low frequency of one observation, sum scores 16, 18 and 20 are presented without a 95% confidence interval. The results for sum scores 16 through 23 were combined to facilitate the interpretation of the results. Figure 4.3b shows the results. The student having the lowest TTM sum score of 16 had a very low l_z statistic. Figures 4.3a and 4.3b show that the mean l_z value stabilized starting from sum score of 25 onward at a mean l_z value just under 0. This result indicates that starting from sum score of 25 onward, the item-score patterns on the low-stakes pre-test were consistent with the proficiency parameters estimated for the high-stakes End of Primary School Test. The low-stakes administration condition of the pre-test did not (or very little at most) influence these item-score vectors. The 95% confidence interval for the mean l_z statistic for TTM sum score 36 was only little wider than that for sum score 25 onward.

Social desirability

Based on the data of 1,484 students (incomplete cases removed), Table 4.2 presents the SD items and their means and standard deviations. The relation-a.



b.

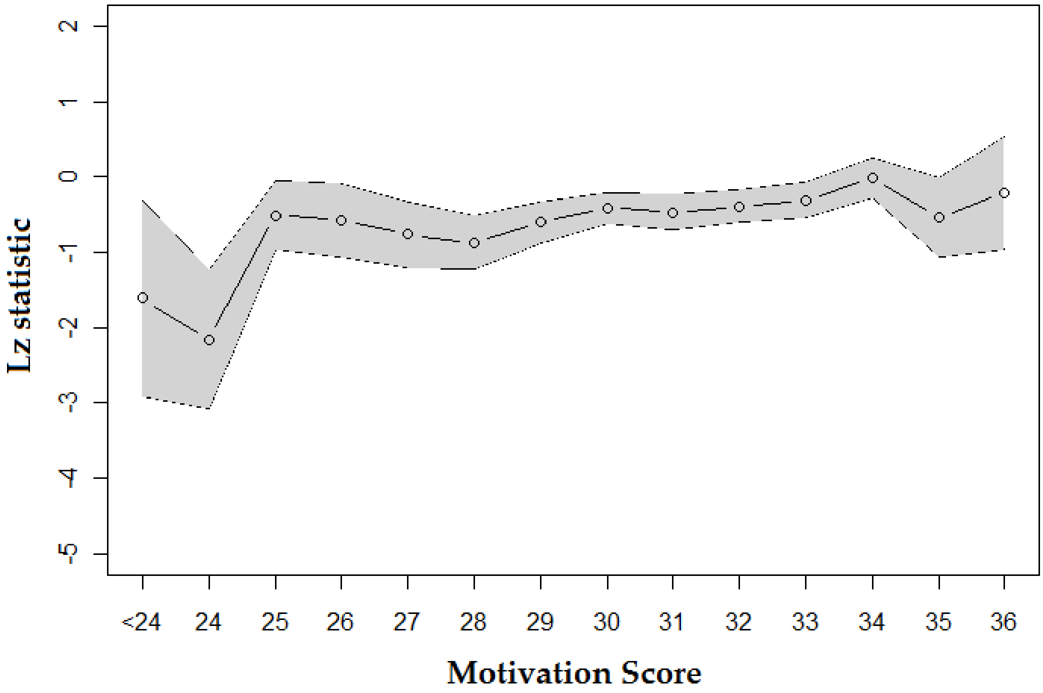


Figure 4.3 *(a) Mean l_z statistic per motivation score, (b) mean l_z statistic per motivation score with the lowest sum scores on the TTM scale combined. The gray area represents the 95% confidence interval for the mean l_z statistic.*

ship between the TTM score and the SD scale score was investigated by means of the Kruskal–Wallis Test. The results revealed a statistically significant difference between the TTM sum score across the seven different SD scores (Group 1, n = 286: Sum score 6; Group 2, n = 259: Sum score 7; Group 3, n = 269: Sum score 8; Group 4, n = 231: Sum score 9; Group 5, n = 196: Sum score 10; Group 6, n = 131: Sum score 11; Group 7, n = 65: Sum score 12), $\chi^2(6, n = 1{,}437) = 92.08$, $p < .001$. The higher SD sum scores, 11 and 12, recorded a higher median sum score on the TTM scale (*Md* = 32) than the SD sum scores 7 through 10 (*Md* = 31) and the SD sum score equal to 6 (*Md* = 30). As the results showed a statistically significant difference between the TTM sum scores across the different SD scores, the analyses were rerun without the highest SD sum score, which was equal to 12. Removing these cases from the analyses did not change the results regarding the relationship of the TTM scale sum score and the posterior probabilities of the mixture Rasch model on the one hand and the $l_z$statistic on the other hand. Therefore, the results of the complete dataset were interpreted.

Discussion

The validity and the reliability of the TTM scale have not been investigated in earlier studies. Consequently, the question that arises is whether the TTM scale is appropriate as a measure of self-reported motivation. A more extensive investigation of the TTM scale is desirable. The reliability (GLB) was appropriate for the type of inference envisaged (Evers et al., 2010). The PCA revealed an internal structure approximately corresponding to results reported in existing literature on TTM (Eklöf, 2006). For example, two of three motivational components that Eklöf found in the development of the TTM Questionnaire (TTM, general attitudes, and performance expectancy) were also found for our TTM scale. The fact that we did not find a "performance expectancy" component might be due to the limited number of items in the TTM scale measuring performance expectancy. Furthermore, our TTM scale was administered to younger children who are probably affected more by "*external motivation*" than older children. The relationship between the TTM sum score and SD was as we expected. Higher SD scores were associated with higher TTM scores. Most likely, this result explains why the 95% confidence intervals found with the highest TTM sum score in Figure 1 and 3 are slightly wider than the confidence intervals found with the lower sum scores. This result was probably due to response tendencies or the influence of SD on the maximum TTM score. We conclude that the TTM sum score can be used in our research as a measure of self-reported motivation.

The relationship between the posterior probability and the TTM sum score did not provide an indication of whether the posterior probabilities of the mixture Rasch model are useful for modeling motivation in low-stakes administration conditions. Even though the mean posterior probabilities increased when the TTM sum score increased, it is not certain whether the two latent classes the mixture Rasch model estimated actually represent "low-stakes" and "high-stakes" response behavior. After all, the correlation between the TTM sum score and the posterior probabilities was low. Furthermore, the mean posterior probability stabilized at approximately .6. If the latent classes truly represented "low-stakes" and "high-stakes" response behavior, the mean posterior probability likely would increase more among the higher TTM sum scores. Possibly, the classes did not represent "low-stakes" and "high-stakes" response behavior, but instead reflected something else. Furthermore, the lower difficulty of items in the class representing "low-stakes" response behavior might indicate that assuming that the student's ability did not differ across latent classes was incorrect. An in-depth analysis of the interpretation of the latent classes is needed. For now, we conclude that the posterior probabilities of the mixture Rasch model have a limited usefulness in modeling motivational differences.

The l_z statistic seemed a more promising approach to model motivational differences. First, the correlation between the l_z statistic and the TTM sum score suggested a stronger relationship. Second, not only did the student having the lowest TTM sum score have the lowest l_z statistic, the mean l_z statistic stabilized just below 0 among the higher TTM sum scores, which was expected. Even though the l_z statistic seems more useful in modeling motivational differences, the results should be interpreted with caution. The parameter estimates on which the l_z statistic is based were estimated in a dataset including highly misfitting item-score vectors. Consequently, the data of a relatively small cluster of students showing extreme response behavior might have influenced the parameter estimates. Therefore it is advisable to only use the l_z statistic as a means for identifying the most extreme cases instead of whole classes displaying "low-stakes" response behavior. Without using an iterative procedure to update the parameter estimates and the l_z statistics, this approach is most likely will fail to identify whole classes as displaying "low-stakes" response behavior.

The results lead us to conclude that the l_z statistic may be particularly useful for modeling motivational differences. However, it would be wise to only use the l_z statistic to identify the most aberrant item-score vectors, especially when students producing these aberrant item-score vectors behave as can be expected under low-stakes and high-stakes administration conditions.

Chapter 5

Using Person-Fit Statistics to Investigate the Effect of Differential Motivation on Educational Test Performance[*]

Abstract

If the stakes in testing are low, students may care little whether their test scores accurately reflect their maximum performance level. Real data studies indicate that the absence of motivation may have a negative effect on a student's test scores and the consistency of a student's responses. This study investigated the difference between responding in low-stakes and high-stakes administration conditions in relation to test performance and response consistency. A response consistency difference occurred more often than a difference in performance in the administration conditions. Students differing on account of both consistency and performance were rare. Scores on a test-taking motivation questionnaire significantly explained variation in (1) the response consistency on the low-stakes tests and (2) the differences in performance on the low-stakes and high-stakes tests. The proportion of explained variance was small.

[*] This chapter has been submitted for publication

Educational tests are used to measure a student’s proficiency on the latent-variable scale. Herein, it is implicitly assumed that the test score accurately reflects a student’s aptitude on the attribute of interest. In terms of performance measurements, if no important personal consequences are associated with the test outcome, students may care little whether their test scores accurately reflect their maximum performance level (Reise & Flannery, 1996; Wise & DeMars, 2005). This phenomenon is called differential motivation (Holland & Wightman, 1982) and refers to the difference in test-taking motivation that exists between high-stakes (i.e., consequential) administration conditions, where students are assumed to pursue maximum performance, and low-stakes (i.e., non-consequential) administration conditions that usually elicit typical rather than maximum performance. The lack of motivation on low-stakes tests is likely to be a concern for assessment professionals who, for example, wish to use data from low-stakes administration conditions for research purposes, to pilot items in low-stakes conditions that are intended for use in high-stakes tests, or to use low-stakes assessments to evaluate the quality of schools (Wise & Kong, 2005).

In the field of educational measurement, accumulated evidence suggests that test-taking motivation is higher in high-stakes administration conditions than in low-stakes administration conditions (Sundre, 1999; Wolf, Smith & Birnbaum, 1995) and that highly motivated students tend to perform better than less well motivated students (Liu, Bridgeman, & Adler, 2012). Wise and DeMars (2005) conducted a meta-analysis of 12 studies and reported 25 effect size statistics, which reflected the mean performance difference between motivated and unmotivated students. All but one of the 25 effect sizes were found to be significant and positive with a mean effect size of 0.59, indicating that on average the mean performance difference between motivated and unmotivated students was more than one-half of a standard deviation. Even though younger students tend to take tests more seriously than older students (Paris, Lawton, Turner, & Roth, 1991), low test-taking motivation and its negative effect on test scores has been found in both elementary (Brown & Walberg, 1993) and higher education contexts (Kiplinger & Linn, 1996; Liu, Bridgeman, & Adler, 2012; O’Neill, Sugrue, and Baker, 1996).

Real data studies found that low test-taking motivation negatively affects test scores in different ways. For example, Wise and Kong (2005) argued that the effort an examinee devotes to an item may vary throughout the test. Inferring rapid-guessing behavior from item response times (i.e., item response times are lower than the time needed to read and ponder the item; Wise, 2007) showed that it is reasonable to believe that some examinees may start motivated but are less motivated from a certain item onwards, which is reflected in their rapid-guessing behavior. Furthermore, Wolf, Smith, and Birnbaum (1995) found evidence that the effect of the stakes that the administration condition has on test performance differs substantially between different types of items. Compared to items that were not mentally taxing, the performance on items that were mentally taxing was affected more by the difference between the stakes of the administration conditions. Interestingly, both examples indicate that differential motivation may not only have a negative effect on a student's test scores but also on the consistency of his or her responses.

Response consistency refers to the degree to which the student's observed item scores equal his expected item scores based on his latent trait value (Conijn, Emons, Van Assen, & Sijtsma, 2011). To illustrate this, failing two relatively easy math items, such as $5 + 4 =$.. and $6 - 3 =$.. is consistent with a relatively low latent trait value because failing both items suggests low math ability. However, passing a relatively difficult math item, such as $18 : 3 =$.., and failing a relatively easy item, such as $5 + 4 =$.., is inconsistent with every latent trait value. The consistency of an individual item-score vector can be investigated by means of person-fit statistics. Person-fit statistics evaluate the fit of an individual's observed item-score vector by comparing the observed item scores to the item scores most likely according to a particular item response theory (IRT) model (Meijer & Sijtsma, 2001). Studies investigating the usefulness of person-fit statistics for analyzing empirical data found some evidence that groups of respondents with known a priori characteristics, such as low test-taking motivation, responded aberrantly (Meijer & Sijtsma, 2001).

Most person-fit methods only allow a binary decision about whether a respondent's item-score vector is aberrant, thereby neglecting a recovery of the many possible mechanisms that caused the aberrance (Meijer & Sijtsma, 2001; Tellegen, 1988). The presence of aberrant item-score vectors in low-stakes data could be due to low test-taking motivation as well as to cheating or alignment errors on the answer sheet. Auxiliary information should be used to investigate the source of aberrance. Additionally, if students' data are available from different test forms measuring the same attribute, a person-fit analysis of the item-score vectors from different test forms may provide information about the psychological processes that explain misfit (Ferrando, 2014).

Developing methods to investigate threats to valid measurements is valuable for measurement practice and research. Gaining knowledge of the effect that the stakes of the administration condition has on the individual item-score vector is therefore desirable. The goal of the current study was to investigate the difference between responding in low-stakes and high-stakes administration conditions in relation to test performance and response consistency. Data from Dutch primary school students who were administered both a math test in a low-stakes condition and a math test in a high-stakes condition were used to answer the following research questions:

1. *Do differences between high-stakes and low-stakes administration conditions exist in relation to the performance and consistency of the individual item-score vector?*

It is expected that students perform better and show more consistent response behavior on tests administered in high-stakes administration conditions than on tests administered in low-stakes administration conditions. The effects of differential motivation on performance and response consistency were investigated using two statistics. First, a Lagrange multiplier- (LM; Glas & Dagohoy, 2007) based person-fit statistic was used to investigate differences between performance on both math tests by assessing whether the proficiency parameter for one student was the same in both tests. Second, the much used global l_z person-fit statistic was used to separately assess a student's consistency in both item-score vectors (Drasgow, Levine, &Williams, 1985). The overlap, or lack thereof, between subsets of students detected by different person-fit statistics was used to assess the degree to which differences between *proficiency* on the two tests coincided with a difference between *consistency* on the two tests. Stability of inconsistency across high-stakes and low-stakes administration conditions would suggest that response inconsistency was due to a stable tendency rather than, for example, a lack of motivation on one measurement occasion due to differential motivation.

2. *Are differences in response behavior found between data from high-stakes and low-stakes administration conditions related to auxiliary information, such as self-reported test-taking motivation, gender, and parental socioeconomic status (SES)?*

To address this research question, a test-taking motivation questionnaire was presented to the students who were administered the low-stakes math tests. If differences in performance and consistency between high-stakes and low-stakes administration conditions are related to gender, parental SES, or self-reported test-taking motivation, these variables can be used to identify students at risk of producing invalid results.

Method

Participants and Data Collection Design

Three different scales were used to collect data: the mathematics scale of the End of Primary School Test 2013 (a high-stakes test administered in the Netherlands at the end of primary education), the pre-test of the mathematics scale of the End of Primary School Test 2014, and a scale measuring test-taking motivation.

Items are usually pre-tested to examine their psychometric properties before including them in a high-stakes test. To pre-test math items intended for use in the End of Primary School Test 2014, a convenience sample of eighth-grade primary school students (N = 9,943; 49.3% male) were presented with different sets of mathematics items in January 2013 (henceforth called 'the pre-test'). The sample was representative of Dutch primary schools according to region, school size, and indicators of parental SES. A random sample of 250 schools was drawn from all schools participating in the pre-test. These schools were asked to administer a questionnaire measuring test-taking motivation directly after the administration of the pre-test. Fifty-one schools attended by 1,199 eighth-grade students (47.5% male), agreed to participate. In February 2013, all schools administering the test-taking motivation questionnaire participated in the End of Primary School Test 2013.

Measures

Pre-test. Students were presented with sets of items varying in number from 30 to 60. In total, 638 multiple choice items were pre-tested. The responses were coded with 0 representing an incorrect answer and 1 representing a correct answer. Since a given pre-test item was administered in more than one item set, the number of observations per item ranged from 25 to 1,785. No performance feedback on the pre-test was provided. Therefore, it was considered that the items were administered in a low-stakes condition.

Test-taking motivation questionnaire. The test-taking motivation questionnaire consisted of 18 items. The construction of the items was inspired by existing scales, such as the Test-Taking Motivation Questionnaire (Eklöf, 2006), the Student Opinion Scale (Thelk, Sundre, Horst, & Finney, 2009), and a subset of items from the self-report questionnaires of the Education Quality Accountability Office (Zerpa, Hackey, Van Barneveld, & Simon, 2011). The

Table 5.1 *Test-taking motivation items*

Item	
1	I enjoy going to school.
2	I enjoy learning math.
3	I am good at math.
4	I think it is important to learn math.
5	I think it is always important to do your best on a test.
6	I did my best on the math items.
7	I answered the math items quickly.
8	I answered the math items seriously.
9	My teacher wants me to do my best on the math items.
10	My parents want me to do my best on the math items.
11	The kids in my class did their best on the math items.
12	I guessed the answer for some of the more difficult items.
13	I'm curious about how many math items I answered correctly.
14	Even with the dull math items, I tried to do my best.
15	I did a good job on the math items.
16	I don't like it when I have to make multiple calculations to answer an item.
17	I could have worked harder on the math items.
18	I could not have done a better job answering the math items.

questionnaire included items specifically assessing test-taking motivation during the pre-test and items assessing test-taking motivation in general. Each item was answered on a 4-point Likert scale (1 = No, 2 = Not so much, 3 = Kind of, 4 = Yes). Table 5.1 shows the English translations of the items.

End of Primary School Test 2013. The End of Primary School Test is administered to students in the final year (i.e., eighth grade) of Dutch primary education. Each year, approximately 85% of all primary schools in the Netherlands, representing the same percentage of students, participate in the test. The test results provide independent advice to primary school teachers, parents, and secondary schools about the most appropriate type of secondary education for a student, thus rendering it important from the perception of all involved. The test is administered in a high-stakes condition, and the secrecy of the items is vital. The End of Primary School Test 2013 contained 60 multiple choice math items, and the responses were coded with 0 representing an incorrect answer and 1 representing a correct answer.

Statistical Analyses

Two-parameter logistic model. The data of the pre-test and the data of the End of Primary School Test 2013 were combined, providing 9,943 item-score vectors containing scores on items administered in a low-stakes condition (pre-test) and different items administered in a high-stakes condition (End of Primary School Test 2013). A two-parameter logistic (2PL) model (Embretson & Reise, 2000) was estimated for this dataset using the software package MIRT (Glas, 2010). In the 2PL model, the probability of a correct response depends on two-item parameters, α_i and β_i, which are interpreted as the item discrimination parameter and the item difficulty parameter of item i, respectively. The 2PL model is given by

$$P(X_i = 1|\theta_j) = \frac{\exp[\alpha_i(\theta_j - \beta_i)]}{1 + \exp[\alpha_i(\theta_j - \beta_i)]},$$

where θ_j is the proficiency parameter for respondent j. The 2PL model must fit the math items of the pre-test and the End of Primary School Test 2013 sufficiently well to allow a meaningful assessment of response consistency relative to the 2PL model. First, the fit of the 2PL model was compared to the fit of the more restrictive Rasch model or 1PL model in which all α_is are equal, and the β_is can vary freely by comparing the log-likelihoods of both models. Second, we assessed the item fit of the 2PL model by evaluating the fit of the item characteristic curve by means of an LM test provided by MIRT. This LM test is performed on each item and is based on creating three different score groups and comparing the observed and expected item scores within the score groups (Glas, 1999).

Lagrange multiplier test for the constancy of the latent variable. Since each student completed both the low-stakes pre-test and the high-stakes End of Primary School Test 2013, differences between the performances on the two math tests might be assessed by comparing the proportion of correct responses for each test for each student. However, because the two tests comprised different items, score differences between the tests could be attributed to differences between the difficulty of the tests, changes in proficiency between the two test administrations, or to differences between the stakes of the administration conditions. Glas and Dagohoy (2007) proposed a person-fit test based on the LM statistic. The LM test was used to test the constancy of the proficiency parameter across subtests and can therefore be used to assess whether the same proficiency estimate can be used to model different subsets of item scores obtained under different administration conditions. A drawback of most person-fit statistics is that their asymptotic distribution is unknown (Nering, 1995) since the derivation of the distribution of the statistics has to account for the uncertainty in the estimated proficiency parameter. The LM test takes the effect of the proficiency estimation into account. The LM test for the constancy of theta was computed by means of MIRT (see Glas and Dagohoy (2007) for a detailed description of the computation of the LM test). An item-score vector was classified as aberrant if the p-value corresponding to the LM test was lower than .05.

The l_z^* statistic. The l_z statistic is a well-known person-fit statistic (Drasgow et al., 1985), which assesses the likelihood of an item-score vector under a specific IRT model. The l_z statistic is assumed to be a standard normal deviate with large negative values that provide evidence of misfit. However, research has shown that the normal approximation to l_z is invalid, thereby yielding a conservative test, particularly for detecting aberrant responses at the lower and higher end of the latent-trait scale (van Krimpen-Stoop & Meijer, 2002). A conservative test implies a loss of power to detect aberrant item-score vectors, which may seriously hamper the usefulness of person-fit analyses of real data. A modified version of l_z denoted as l_z^* has been proposed (Snijders, 2001, also, see Van Krimpen-Stoop & Meijer, 1999) for which a valid asymptotic theoretical sampling distribution was derived. The computation of l_z^*, given the 2PL model and using weighted maximum likelihood (WLE; Warm, 1989) estimators for the proficiency parameters, is presented in Appendix A.

For each student, the l_z^* statistic was computed using dedicated software on two different item-score vectors, which were the item-score vectors containing scores on the (low-stakes) pre-test and the item-score vector containing scores on the (high-stakes) End of Primary School Test 2013. The item parameters used for computing l_z^*low-stakes and l_z^*high-stakes resulted from estimating the 2PL model on the complete dataset. The proficiency parameters used for computing both l_z^*low-stakes and l_z^*high-stakes were estimated from the item-score vector from the pre-test or the item-score vector from the End of Primary School test, respectively. Since large negative values of l_z^* provide evidence of misfit, an item-score vector was classified as aberrant if l_z^* was lower than -1.64 (i.e., α = .05, one-tailed).

Variation in person misfit

After having estimated the 2PL model and the person-fit statistics from the 9,943 item-score vectors, the overlap between the subsets of students detected by different person-fit statistics was assessed for the students who completed the test-taking motivation questionnaire. The different types of person misfit investigated were represented in a Venn diagram to investigate the overlap between the subsets of students detected by the LM statistic, l_z^*high-stakes and l_z^*low-stakes.

Explaining person misfit

Since students are nested within schools, a multilevel regression model, which takes into account the variance within and between schools, was estimated to investigate whether the person-fit statistics were related to the explanatory variables. In these analyses, the dependent variable of person fit is treated as a continuous variable. The intraclass correlation (ICC) (Snijders & Bosker, 1999, pp. 16-22) was used to assess which part of the total variance could be attributed to the school level. The explanatory variables included gender, parental SES, and the score on the test-taking motivation questionnaire. Since the test-taking motivation questionnaire contained items specifically related to the administration of the pre-test, it might seem odd to include test-taking motivation as an explanatory variable for l_z^*high-stakes. However, even though a large effect was not expected, it was interesting to inspect the extent to which the score on the test-taking motivation questionnaire was better at explaining l_z^*low-stakes than l_z^*high-stakes. A variable indicating parental SES consisted of three categories with one being high on parental SES and three being low on parental SES. After recoding the contra-indicative items of the test-taking motivation questionnaire (i.e., items 12, 16, and 17), the greatest lower bound (GLB) found from a factor analysis was calculated as a reliability estimate using the psych-package (Revelle, 2014) from the statistical program R (R Core Team, 2013).

Results

Model fit

A chi-square test was used to assess the fit of the 2PL model relative to the more restrictive Rasch model. It was concluded that the 2PL model fit the data significantly better than the Rasch model (χ^2 = 8,832.23, *df* = 698, *p* <.001).

The fit of the item characteristic curves was assessed using the LM test. Glas (2010) argued that with large sample sizes, the absolute difference between the observed and the expected item scores for each score group is more informative about model violation than the significance level. We decided to inspect items which showed significant deviation from the expected item characteristic curve and for which the absolute difference exceeded .05.

Large deviations from the expected item characteristic curve were found in two items on the pre-test. For one item, we were unable to provide an explanation for the deviation, but for the other item, it appeared that a graph needed to correctly answer the item was not printed in one of the pre-test booklets, which rendered answering the item correctly impossible without guessing. We decided to remove the item in this particular pre-test booklet for further analyses. After re-estimating the 2PL, the number of misfitting items per pre-test booklet was approximately five percent, which we considered an acceptable model fit based on the Type I error rate expected.

Variation in person misfit

Figure 5.1 shows a Venn diagram with the number of detected students and the overlap between the subsets of students detected by each statistic: the LM person-fit statistic, l_z^* estimated on pre-test items (l_z^*low-stakes), and l_z^* estimated on the End of Primary School Test 2013 (l_z^*high-stakes). The detection rate of the LM statistic was higher (10.7%) than the detection rate of l_z^*high-stakes (7.7%) and l_z^*low-stakes (5.9%). It should be noted that since a base rate of aberrant item-score vectors is unknown, the detection rate of the different person-fit statistics does not provide us with information about the performance of the person-fit statistics. Interestingly, the overlap between the subsets of students detected by l_z^*low-stakes and l_z^*high-stakes was small, indicating that response inconsistency identified by l_z^* cannot be considered a stable tendency across different test forms. The overlap between the subsets of students detected by the LM statistic on the one hand and l_z^*high-stakes and l_z^*low-stakes on the other hand was also small. No student was classified as aberrant by all three person-fit statistics. These results suggest that a difference between the performance on high-stakes and low-stakes tests usually does not coincide with (1) the response inconsistency on either the high-stakes or the low-stakes tests; and (2) the transitions between the response inconsistency on the tests (e.g., showing inconsistent responding on the low-stakes test but consistent responding on the high-stakes test).

Explaining person misfit

We estimated a multilevel regression model to explain person misfit, as measured by the LM test, the l_z^*high-stakes test, or the l_z^*low-stakes test. This regression model enabled us to take both the variance of the person-fit statistics within and between schools into account. Gender, an indicator of parental SES, and the score on the test-taking motivation questionnaire were included as level-1 independent variables. The GLB for the total test-taking motivation

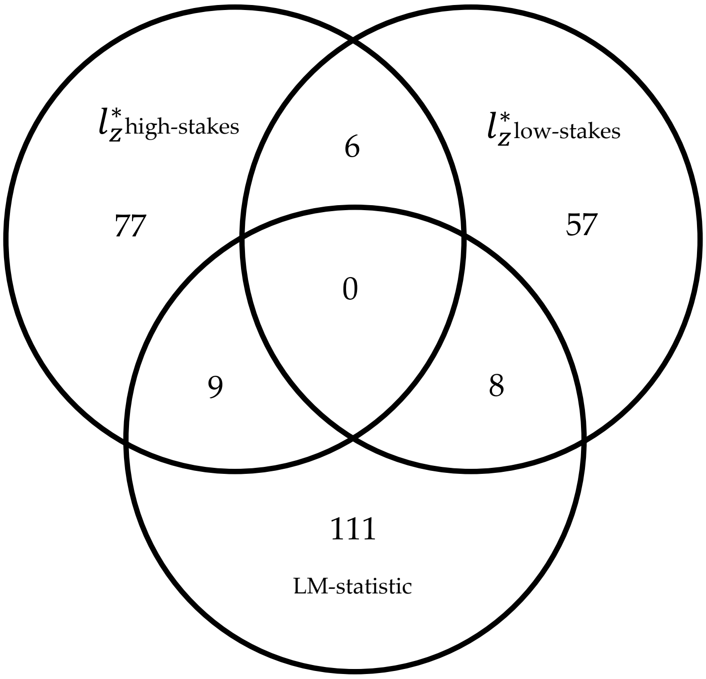


Figure 5.1 *Venn diagram showing the overlap between the subsets of students detected by* l_z^**high-stakes*, l_z^**low-stakes and the LM statistic (N = 1,199).*

questionnaire was equal to .77. This value is suggestive of a reliability that allows less important decisions about individuals (Evers et al., 2010). The regression coefficients, the model fit indices and the variance components of the null model, and the model with predictors (i.e., the full model) are presented in Table 5.2.

Based on the relative fit statistics known as the -2 log likelihood (LL), the Akaike information criterion (AIC), and the Bayesian information criterion (BIC; Singer & Willett, 2003, pp. 119-122), we concluded that for each person-fit statistic, the full model had to be preferred over the null model. For the LM statistic, only the score on the test-taking motivation questionnaire was a significant predictor. For the l_z^*low-stakes, both the score on the test-taking motivation questionnaire and one dummy variable (comparing category 1 and 3) of parental SES were significant. The different categories of parental SES differed significantly on l_z^*low-stakes (p < .05). None of the predictors were significantly related to l_z^*high-stakes. The student-level explained proportion of variance, R_1^2, was small for each person-fit statistic. For the LM statistic and the l_z^*low-stakes statistic, R_1^2 was negative. The 95% confidence intervals for the variance of the individual-level residuals (σ^2) and the variance of the school-level residuals (τ^2) for the null model and the full model indicated that the

Table 5.2 *Estimated regression coefficients, model-fit indices and variance components for each person-fit statistic*

	LM		l_z^*low-stakes		l_z^*high-stakes	
Variable	*Null model*	*Full model*	*Null model*	*Full model*	*Null model*	*Full model*
Intercept	1.49	4.01	0.38	-1.20	-0.12	-0.58
Male	-	-0.09	-	-0.15	-	-0.08
SES1	-	-0.41	-	0.43*	-	0.11
SES2	-	-0.44	-	0.24	-	0.33
TTM	-	-0.03*	-	0.02**	-	0.01
Model Fit						
-2LL	5,608.17	4,598.67	3,849.47	3,111.97	3,522.20	2,819.75
AIC	5,612.17	4,612.67	3,853.47	3,125.97	3,526.20	2,833.75
BIC	5,622.35	4,646.76	3,863.66	3,160.10	3,536.39	2,867.88
Variance						
σ^2	6.09	6.80	1.36	1.39	1.06	1.04
τ^2	0.20	0.18	0.11	0.12	0.04	0.04
ICC	0.03	0.03	0.07	0.08	0.04	0.04
R_1^2	-	-0.09	-	-0.02	-	0.02

Note. N = 1,199, SES1= parental SES category 1 compared with parental SES category 3; SES2 = parental SES category 2 compared with parental SES category 3; TTM = score on the test-taking motivation questionnaire.
*p < .05. **p < .01.

negative R_1^2 value was most likely due to a chance fluctuation rather than the misspecification of the model (i.e., random regression coefficients as opposed to fixed regression coefficients; for a more detailed discussion on the negative R_1^2, see Snijders and Bosker (1999, pp. 99-104)).

Discussion

We investigated whether the differences between high-stakes and low-stakes administration conditions existed in the context of test performance and the consistency of the individual item-score vector. It was concluded that differences in performance and consistency existed between the administration conditions. However, a difference in consistency between the administration conditions (i.e., showing inconsistent responding on the low-stakes test but consistent responding on the high-stakes test or vice versa) occurred more often than a difference in performance. Students showing both a difference in performance and a difference in consistency were rare.

As mentioned earlier, there are different explanations on how low test-taking motivation negatively affects test scores. Examples include students who are less motivated by items that are considered mentally taxing or those who lose motivation near the end of the test. Given that the position of the item or the degree to which it is considered mentally taxing is not perfectly related to the difficulty of the item, in the situations described in both examples, one expects to find a difference between high-stakes and low-stakes tests with respect to performance *and* consistency. However, we did not find significant overlap between the subsets of students who were classified as aberrant with respect to the consistency of the item-score vector in both high-stakes and low-stakes administration conditions. The question arises as to how students approach a low-stakes test. Theoretically, the degree to which an item is mentally taxing (i.e., the mental effort needed to reach a correct answer) differs from the difficulty of this item (i.e., the p-value) (Wolf, Smith, & Birnbaum, 1995). Despite this theoretical difference, it remains unclear whether students distinguish between mental workload and item difficulty in practice. If not, then only a shift in performance is likely. Furthermore, even though the level of difficulty of the items usually varies throughout the test form, it is also common to start with relatively easy items and then slowly increase the difficulty of the items. However, this is not done monotonically, that is, as items become more difficult, they are sometimes followed by easier items to keep students motivated. Thus, even though the relationship between item position and item difficulty is not perfect, it is not unrelated either. This relationship suggests that if students are indeed less motivated toward the end of the low-stakes test, the amount of inconsistency would still not be large. It is questionable whether person-fit statistics, such as l_z^*, have sufficient power to detect small effects unless tests became unrealistically long (Emons, Sijtsma, & Meijer, 2005; Meijer & Sijtsma, 2001; Reise & Due, 1991).

The lack of overlap between the subsets of students classified as aberrant with respect to the consistency of the item-score vector in both the high-stakes and low-stakes administration conditions leads us to believe that person misfit is not a stable tendency. This result contradicts that of Woods, Oltmanns, and Turkheimer (2008) who found positive correlations, some of which substantial, across five temperament and trait scales of the Schedule for Nonadaptive and Adaptive Personality (Clark, 1996). A possible explanation for these contradictory findings is that person misfit in non-cognitive and cognitive measurements results from different psychological processes. This explanation is supported by Schmitt et al. (1999) who found that l_z values estimated from different cognitive ability domains were uncorrelated whereas l_z values estimated from different personality domains were moderately correlated. It is likely that as opposed to the cognitive measurement, misfit in the non-cognitive measurement is influenced more by specific, stable response styles, such as faking a particular personality characteristic well.

After having assessed the different types of misfit between and within high-stakes and low-stakes administration conditions, we tried to explain person misfit by regressing person-fit statistics on explanatory variables. Presumably relevant covariates, such as a score on a test-taking motivation questionnaire, significantly explained variation within the l_z^*low-stakes and the LM statistic. However, the proportion of explained variance was very small.

The studies devoted to explaining person-misfit provided us with mixed results. Schmitt et al. (1999) found significant correlates of person-misfit in non-cognitive measurements (e.g., conscientiousness, gender, and test-taking motivation). Conijn, Emons, Van Assen, Pedersen, and Sijtsma (2013) also found significant covariates of person-misfit in non-cognitive measurements (e.g., conscientiousness and psychopathology), but they only explained small proportions of variation in person misfit. Conijn et al. (2013) suggested several explanations for the low explanatory power in person-fit research. For example, different types of model misfit may be related to different explanatory variables. Consequently, a single regression model does not sufficiently explain person misfit, and different regression models for different types of person misfit may be needed.

Appendix A

Snijders (2001) showed that the asymptotic distribution of

$$l_z^* = \frac{l - E(l) + c_K(\hat{\theta})r_0(\hat{\theta})}{\sqrt{N}\tau_K(\hat{\theta})} \tag{5.1}$$

is standard normal when using the 2PL model and the weighted likelihood estimator (Warm, 1989) where l denotes the unstandardized likelihood of the item-score vector, $E(l)$ denotes the expected likelihood, and K denotes the total number of items. The quantities in Equation 5.1 are given by:

$$l = \sum_{i=1}^{K} \{X_i \ln P_i(\hat{\theta}) + (1 - X_i) \ln[1 - P_i(\hat{\theta})]\}$$

$$E(l) = \sum_{i=1}^{K} \{P_i(\hat{\theta}) \ln[P_i(\hat{\theta})] + [1 - P_i(\hat{\theta})] \ln[1 - P_i(\hat{\theta})]\}$$

$$c_K(\hat{\theta}) = \frac{\sum_{i=1}^{K} a_i\,(\hat{\theta} - b_i)P_i'(\hat{\theta})}{\sum_{i=1}^{K} a_i\,P_i'(\hat{\theta})}$$

$$r_0(\hat{\theta}) = \frac{J\hat{\theta}}{2I\hat{\theta}}$$

$$\tau_K^2\hat{\theta} = \frac{1}{K}\sum_{i=1}^{K} [a_i(\hat{\theta} - b_i) - a_i c_K(\hat{\theta})]^2 P_i(\hat{\theta})[1 - P_i(\hat{\theta})]$$

$$\boldsymbol{I}(\hat{\theta}) = \sum_{i=1}^{K} \frac{P_i'^2(\hat{\theta})}{P_i(\hat{\theta})[1 - P_i(\hat{\theta})]}$$

$$\boldsymbol{J}(\hat{\theta}) = \sum_{i=1}^{K} \frac{P_i'(\hat{\theta})P_i''(\hat{\theta})}{P_i(\hat{\theta})[1 - P_i(\hat{\theta})]}.$$

P_i' and P_i'' are the first and second derivatives of P_i, respectively, with respect to θ, and I and J are the information and Jacobian, respectively.

Chapter 6

Epilogue

Understanding the complexity of the psychological process underlying differential motivation is essential for attempting to model it. A prominent theory regarding motivation is the expectancy-value theory for achievement performance proposed by Wigfield (1994), which argues that achievement performance depends on students' expectancies of succeeding a task and the value they assign to succeeding. Eccles et al. (1983) defined attainment value as the subjectively assigned importance of doing well on a task. In relation to differential motivation, we assumed that the attainment value of passing an item is lower in low-stakes conditions than high-stakes conditions, thus lowering a student's subjective value assigned to passing an item, and thus lowering a student's achievement performance.

The goal of this thesis was to investigate the effect of differential motivation on linking and the possibility to statistically model this effect. Substantial evidence shows that students respond differently to tests administered in low-stakes conditions than to tests administered in high-stakes conditions. The main question was not, therefore, whether an effect of differential motivation existed, but how it would affect the individual item-score vector, the test results and, more specifically for this thesis, the linking result. In this Epilogue, I will reflect on the decisions made in operationalizing and modeling differential motivation. These reflections might serve as input for avenues of future research.

The effect of differential motivation was operationalized by creating a heterogeneous population. This was accomplished by dividing a population assumed to be homogeneous (i.e., the administration condition was high-stakes) into two subgroups, the first containing students who did not respond differentially between administration conditions and the second who responded differentially. Following the mixture Rasch modeling approach, we attempted to model these subgroups by identifying two latent classes. Using the person-fit approach, these subgroups were identified by making a binary decision between fit (i.e., students who performed in a 'high-stakes manner') and misfit (i.e., students performing in a 'low-stakes matter'). Different reasons exist for assuming that the use of (1) a more comprehensive operationalization of differential motivation or (2) a different modeling strategy might be useful for further investigating and modeling differential motivation.

Operationalizing differential motivation. One of the first concepts I encountered, which suggested that the psychological process underlying the effect of differential motivation was more complex than expected, was test anxiety. In light of the expectancy-value theory, it has been shown that students who highly value success (e.g., passing an item) but expect to do poorly, report higher levels of test anxiety (Selkirk, Bouchy, & Eccles, 2011). Regarding the difference between high-stakes and low-stakes administration conditions, Zohar (1988) provided evidence that disposition to anxiety and being tested in a high-stakes administration condition contribute to increased test anxiety. Additionally, a meta-analysis performed by Hembree (1988) showed that test anxiety is related to lower performance. According to this line of reasoning, for students vulnerable to test anxiety, performance on a test administered in a high-stakes administration condition is expected to be lower than performance in a low-stakes administration condition.

Even though the findings from Chapters 1 through 5 suggest that, overall, students perform better on a test administered in a high-stakes administration condition than a low-stakes administration condition, at the individual level, the effect may be reversed because of, for example, test anxiety. The inter-individual differences with respect to responding differently between high-stakes and low-stakes administration conditions are likely to be substantial. This expectation is supported by the different findings in the literature on how differential motivation affects the individual item-score vector. Examples include students who are less motivated by items considered mentally taxing and students who lose motivation near the end of the test.

Additional support for the expectation can be found in the low correlations presented in Chapter 5 between the score on the TTM on the one hand and the posterior probabilities of the mixture Rasch model and the person-fit results on the other hand. The substantial inter-individual differences between responding in high-stakes and low-stakes administration conditions make it hard to believe that these differences can be modeled sufficiently precise by operationalizing differential motivation in just two subgroups. Even though the subgroup of students who do not respond differently between administration conditions is likely to be homogeneous, the subgroup of students who do respond differently is most likely not homogeneous.

Another assumption underlying our operationalization, worthy of discussion, is that students belong to a homogeneous population when tests were only administered in high-stakes conditions. Application of the mixture Rasch model to the internal anchor (in Chapter 2) did not yield evidence for differently motivated subgroups. However, this lack of evidence does not rule out other psychological processes, which might add to variation in responding within administration conditions. In Chapter 3, for simulated data, I found that both the effect size and the number of students showing low-stakes responding needed to be substantial for the mixture Rasch model to identify the simulated latent classes. A possible explanation might be that the variation between administration conditions needs to be sufficiently large compared to the variation within administration conditions before the model can correctly identify the latent classes.

Modeling differential motivation. Identification of latent classes underlying the data only has meaning relative to the operationalization of the construct and the statistical method used (Bouwmeester & Sijtsma, 2007; Bouwmeester, Vermunt, & Sijtsma, 2007). To put it sharply, the mixture Rasch model can only be interpreted meaningfully to the extent that the phenomenon picked up from the data truly reflects the operationalization of responding differently. In addition, for a person-fit approach, a meaningful identification of subgroups, whose item-score vectors are either consistent or inconsistent with a specific IRT model, depends on the operationalization of the construct and the person-fit method used. The results presented in this thesis should therefore be evaluated relative to the mixture Rasch model or the person-fit methods used.

Following the mixture Rasch modeling approach, it was attempted to model differential motivation by imposing constraints on the latent classes.

Using the mixture Rasch model, we assumed that the latent classes only differ with respect to item difficulty parameters, where the item difficulty parameters are higher for the class representing low-stakes responding than the class representing high-stakes responding. However, this idea might be too simple, because allowing classes to differ with respect to item discrimination parameters might improve modeling the differences in responding between administration conditions. The following line of reasoning provides support for allowing item discrimination to vary between classes. A higher attainment value in high-stakes administration conditions will only result in a difference in performance for those students who are proficient enough to pass the item. For the item characteristic curves (ICCs) in the different latent classes, this means that below a particular proficiency value, they overlap, and above this value, the ICC of the latent class showing low-stakes responding increases more slowly than the ICC of the latent class showing high-stakes responding. ICCs for the latent class representing high-stakes responding are expected to be steeper than ICCs for the latent class representing low-stakes responding when the relationship between the probability of passing an item and proficiency is weaker for the latent class representing low-stakes responding. Empirical research has to provide support for this hypothesis.

References

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of mathematical functions*. New York: Dover Publications.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington DC: American Council of Education.

Bartlett, M. S. (1954). A note on the multiplying factors for various chi square approximations. *Journal of the Royal Statistical Society, 16* (Series B), 296–298.

Baxter, S. D., Smith, A. F., Litaker, M. S., Baglio, M. L., Guinn, C. H., & Shaffer, N. M. (2004). Children's social desirability and dietary reports. *Journal of Nutrition Educational and Behaviour*, *36*, 84–89.

Béguin, A. A. (2000). *Robustness of equating high-stakes tests*. (Doctoral dissertation). Twente University, Enschede, The Netherlands.

Béguin, A. A. (2005). *Bayesian IRT equating with correction for unmotivated respondents on the anchor-test*. (Paper presented at the IMPS 2005 in Tilburg, the Netherlands).

Béguin, A.A. (2008). *Application of Mixed IRT models in IRT linking: combining high-stakes tests with a low-stakes anchor.* (Paper presented at the International Meeting of the Psychometric Society, Durham, NC).

Béguin, A. A., & Hanson, B. A. (2001). *Effect of noncompensatory multidimensionality on seperate and concurrent estimation in IRT observed score equating* (Report No. 01-02). Retrieved from Psychometric Research Centre web site: http://www.cito.nl/~/media/citonl/Files/Onderzoek%20en%20wetenschap/cito mrd report 2001 02.ashx

Béguin, A. A., & Maan, A. (2007). *IRT linking of high-stakes tests with a low-stakes anchor*. (paper presented at the 2007 Annual National Council of Measurement in Education (NCME) Meeting, April 10-12, Chicago).

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331-48.

Bouwmeester, S., & Sijtsma, K. (2007). Latent class modeling of phases in the development of transitive reasoning. *Multivariate Behavioral Research, 42*, 457-480.

Bouwmeester, S., Vermunt, J. K., & Sijtsma, K. (2007). Development and individual differences in transitive reasoning: A fuzzy trace theory approach. *Developmental Review, 27*, 41-74.

Brown, S. M., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *The Journal of Educational Research, 86*, 133-136.

Clark, L. (1996). *Schedule for nonadaptive and adaptive personality (SNAP). Manual for administration, scoring and interpretation*. Minneapolis: University of Minnesota Press.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Conijn, J. M., Emons, W. H. M, Van Assen, M. A. L. M, Pedersen, S. S., & Sijtsma, K. (2013). Explanatory, multilevel person-fit analysis of response consistency on the Spielberger State-Trait Anxiety Inventory. *Multivariate Behavioral Research, 48*, 692-718.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225-244.

Davey, T., Nering, M. L., & Thompson, T. (1997). Realistic simulation of item response data. ACT Research Report Series (pp 97-104). Retrieved from: https://www.act.org/research/researchers/reports/pdf/ACT_RR97-04.pdf

Dinero, T. E., & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement, 1*(4), 581-592.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

Eccles, J., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J., and Midgley, C. (1983). Expectancies, values and academic behaviors. In Spence, J. T. (ed.), *Achievement and Achievement Motives*, W. H. Freeman, San Francisco.

Efron, B., & Tibshirani, R.J. (1993). Confidence intervals based on bootstrap percentiles. In *An Introduction to the Bootstrap* (pp. 168-177). Boca Raton, FL: Chapman & Hall.

Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement, 66*, 643–565.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Emons, W. H. M. (1998). Nonequivalent groups IRT observed-score equating: Its applicability and appropriateness for the Swedish Scholastic Aptitude Test. Twente University.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local and graphical person-fit analysis using person response functions. *Psychological Methods*, *10*, 101-119.

Evers, A., Lucassen, W., Meijer, R., Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de Kwaliteit van Tests (geheel herziene versie)* [COTAN Rating system for test quality (completely revised edition)]. Amsterdam: NIP.

Ferrando, P. J. (2014). A comprehensive approach for assessing person fit with test–retest data. *Educational and Psychological Measurement*. doi: 0013164413518555.

Forsyth, R., Saisangjan, U., & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement*, *5*(2), 175-186.

Glas, C. A. W. (1989). *Estimating and testing Rasch models*. Ph.D. dissertation, University of Twente.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, *64*, 273-294.

Glas, C. A. W. (2010). Multidimensional Item Response Theory (MIRT) [Computer software and Manual]. Enschede, the Netherlands: University of Twente. Retrieved from: http://www.utwente.nl/gw/omd/Medewerkers/medewerkers/glas/

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, *72*, 159-180.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test-anxiety. *Review of Educational Research*, *58*, 47-77.

Holland, P. W., & Rubin, D. R. (Eds.). (1982). *Test Equating*. New York: Academic Press.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Wightman, L. E. (1982). Section pre-equating: A preliminary investigation. In P. W. Holland & D. R. Rubin (Eds.), *Test Equating* (pp. 271-297). New York: Academic Press.

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika*, *39*, 31–36.

Kiplinger, V. L., & Linn, R. L. (1996). Raising the stakes of test administration: The impact on student performance on the National Assessment of Educational Progress. *Educational Assessment*, *3*, 111-133.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with non-random groups. *Journal of Educational Measurement, 22*, 197-206.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. (2nd ed.). New York, NY: Springer Verlag.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement*, *16*, 878.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education motivation matters. *Educational Researcher*, *41*(9), 352-362.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*, 453-461.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods, Instruments and Computers*, *38*, 88–91.

Maier, M. H. (1993). Military aptitude testing: The past fifty years (DMCM Technical Report 93-700). Montery, CA: Defence Manpower Data Center.

Mair, P., Hatzinger, R., & Maier, M. (2010). *eRm: Extended Rasch Modeling*. Retrieved from
http: //CRAN.R-project.org/package=eRm.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review and new developments. *Applied Measurement in Education, 8,* 261–272.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied*

Psychological Measurement, *25*, 107-135.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195-215.

Mittelhaëuser, M., Béguin, A. A., & Sijtsma, K. (2011). *Comparing the effectiveness of different linking designs: The internal anchor versus the external anchor and pre-test data* (Report No. 11-01). Retrieved from Psychometric Research Centre Web site: http://www.cito.nl/~/media/cito_nl/Files/Onderzoek%20en%20wetenschap/cito_mrd_report_2011_01.ashx

Mittelhaëuser, M., Béguin, A. A., & Sijtsma, K. (2013). Modeling differences in test-taking motivation: Exploring the usefulness of the mixture Rasch model and person-fit statistics. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.). *New developments in quantitative psychology* (pp. 357-370). New York: Springer.

Mittelhaëuser, M., Béguin, A. A., & Sijtsma, K. (in press). Selecting a data collection design for linking in educational measurement: Taking differential motivation into account. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & W.-C. Wang (Eds.), *New Developments in quantitative psychology: Presentations from the 78th Annual Psychometric Society Meeting*. New York: Springer.

Nering, M. L. (1995). The distribution of person-fit statistics using true and estimated person parameters. *Applied Psychological Measurement*, *19*, 121-129.

O'Neill, H. F., Sugrue, B., & Baker, E. L. (1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, *3*, 135-157.

Paris, S. G., Lawton, T. A., Turner, J. C., & Roth, J. L. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher*, *20*, 12-20.

R Development Core Team. (2010). R: A Language and Environment for Statistical Computing. [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M. D. (2009). *Multidimensional Item Response Theory Models*. New York, NY: Springer Verlag.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*, *15*, 217-226.

Reise, S. P., & Flannery, Wm. P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education*, *9*, 9-26.

Revelle, W. (2014). Psych: Procedures for personality and psychological research. Illinois, USA. Retrieved from https://CRAN.R-project.org/package=psych Version = 1.4.2.

Rost, J. (1997). Logistic Mixture Models. In: W. van der Linden & R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 449-463).

Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324-332). Münster, New York: Waxmann.

Scheerens, J., Glas, C., & Thomas, S. M. (2003). *Educational evaluation, assessment and monitoring: A systematic approach*. Lisse, The Netherlands: Swets & Zeitlinger.

Selkirk, L. C., Bouchey, H. A., & Eccles, J. S. (2011). Interactions among domain-specific expectancies, values, and gender: Predictors of test anxiety during early adolescence. *The Journal of Early Adolescence*, *31*(3), 361-389.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*, 591-611.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person-fit and effect of person-fit on test validity. *Applied Psychological Measurement*, *23*, 41-53.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford University Press.

Snijders, T. A. B. (2001). Asymptotic distribution of person fit statistics with estimated person parameters. *Psychometrika*, *66*, 331-342.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA, Sage.

Sundre, D. L. (1999). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the annual meeting of the American Educational Research Association, Montreal (ERIC Document Reproduction Service No. ED432588).

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality*, *56*, 622-663.

Thelk, A., Sundre, D. L., Horst, J. S., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale (SOS) to make valid inferences about student performance. *Journal of General Education, 58*, 129–151.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

Van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, *23*, 327-345.

Van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2002). Detection of person misfit in computarized adaptive tests with polytomous items. *Applied Psychological Measurement*, *26*, 164-180.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *One-parameter logistic model (OPLM).* Arnhem: Cito, National Institute for Educational Measurement.

Von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika*, *78*, 605-623.

Von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer.

Von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: an extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*, 389-406.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30,* 1–21.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427-450.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, *6*, 49-78.

Wise, S. L. (2007). *Examinee effort and test score validity.* Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, Connecticut.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1-17.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163-183.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety and test performance. *Applied Measurement in Education*, *8*, 227-242.

Wolf, L. F., & Smith, J. K., & Birnbaum, M. E. (1995). The consequence of performance, test, motivation, and mentally taxing. *Applied Measurement in Education*, *8*, 341-351. Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: Mesa Press.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2008). Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychological Assessment*, *20*, 159-168.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: Mesa Press.

Yamamoto, K. & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent traits and latent class models in the social sciences* (pp. 89-98). New York: Waxmann.

Zeng, L., & Kolen, M. J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement*, *19*, 231-240.

Zerpa, C., Hachey, K., van Barneveld, C., & Simon, M. (2011). Modeling student motivation and students' ability estimates from a large-scale assessment of mathematics. SAGE open. doi: 10.1177/2158244011421803.

Zohar, D. (1988). An additive model of test anxiety: role of exam-specific expectations. *Journal of Educational Psychology*, *90*, 330-340.

Summary

In educational measurement, multiple test forms are often constructed to measure the same construct. Linking procedures are used to disentangle differences between test forms with respect to difficulty and differences between student groups with respect to proficiency, so that scores for different test forms can be used interchangeably. Students' differential motivation can be considered a confounding variable when choosing a data collection design for linking. Differential motivation refers to the difference in test-taking motivation that exists between high-stakes and low-stakes administration conditions (Holland & Wightman, 1982). In a high-stakes administration condition, a student is expected to work harder and strive for maximum performance, whereas a low-stakes administration condition does not challenge students explicitly, and thus may elicit typical, rather than maximum, performance. In this thesis, we first discuss the suitability of different data collection designs and the way they are typically implemented in practice. In Chapters 2 through 5, we investigated the suitability of a mixture Rasch model and person-fit methods to model differential motivation. Constraints on the mixture Rasch model should help identify the latent classes in such a way that one latent class represents high-stakes responding while the other represents low-stakes responding.

In Chapter 2, we used data from a Dutch testing program to investigate whether the differences between estimated proficiency distributions obtained from two operational tests differed between data collection designs with anchor items administered in low-stakes conditions on the one hand and data collection designs with anchor items administered in high-stakes conditions on the other hand. The external anchor design was concluded to be more robust

against the effect of differential motivation than the pre-test design. Specifically, using the pre-test design to link the operational tests resulted in a substantial overestimation of the difference between the estimated mean proficiency of the populations administered the operational tests. The effect of differential motivation in the pre-test design can be controlled for by using a constrained mixture Rasch model to link the operational tests. Removing items displaying DIF between high-stakes and low-stakes administration conditions does not improve the linking result.

The purpose of Chapter 3 was to investigate under which conditions simulated differential motivation between the stakes of operational tests and anchor items produces an invalid linking result when the Rasch model was used to link the operational tests. This was done for an external anchor design and a variation of a pre-test design. Additionally, the question whether a constrained mixture Rasch model can identify the simulated latent classes was also investigated. The results indicate that for an external anchor design, the Rasch linking result is only biased when the motivation level differs between the subpopulations to which the anchor items are presented. We found that the constrained mixture Rasch model did not identify the simulated latent classes representing low-stakes and high-stakes responding. When a pre-test design was used to link the operational tests by means of a Rasch model, in each condition bias in the linking result was found. The amount of bias increased as the percentage of students showing low-stakes responding on the anchor items increased. The constrained mixture Rasch model only identified the simulated latent classes representing low-stakes and high-stakes responding under a limited number of conditions.

In Chapter 4, we explored the extent to which a constrained mixture Rasch model and the l_z person-fit statistic can be used to model motivational differences in data obtained from a low-stakes administration condition. We investigated the usefulness of the mixture modeling strategy in a sample of primary-school students (N = 1,512) by comparing the posterior probabilities of the constrained mixture Rasch model and students' self-reported motivation. Furthermore, we investigated the relationship between students' self-reported motivation and the l_z person-fit statistic. The results led us to conclude that compared to the posterior probabilities of the constrained mixture Rasch model, the l_z person-fit statistic seems a more promising approach to model motivational differences.

Real-data studies indicate that the absence of test-taking motivation may have a negative effect on a student's test score and the consistency of a student's responses. The focus of Chapter 5 on the difference between responding in low-stakes and high-stakes administration conditions in relation to test performance and response consistency showed that a response consistency difference occurred more often than a difference in performance in the administration conditions. Students differing on account of both consistency and performance were rare. Scores on a test-taking motivation questionnaire significantly explained variation in (1) the response consistency on the low-stakes test and (2) the differences in performance on the low-stakes and high-stakes tests. However, the proportion of explained variance was small.

In Chapter 6, reasons were discussed for assuming that the use of (1) a more comprehensive operationalization of differential motivation and (2) a different modeling strategy might be useful for further investigating and modeling differential motivation.

Samenvatting (Summary in Dutch)

Van toetsen worden vaak meerdere versies geconstrueerd die hetzelfde construct beogen te meten. Linkprocedures kunnen gebruikt worden om de verschillen tussen de moeilijkheid van de toetsen en de verschillen tussen de vaardigheden van de leerlingen te onderscheiden. Na correctie via de linkprocedure kunnen scores over verschillende toetsen vergeleken worden. Wanneer we een data design kiezen om te linken, moeten we rekening houden met differentiële motivatie als potentieel storende variabele. De term differentiële motivatie refereert naar het verschil in motivatie tussen toetsen afgenomen in high-stakes afname condities en toetsen afgenomen in low-stakes afname condities (Holland & Wightman, 1982). In een high-stakes afname conditie verwachten we dat een leerling hard werkt en streeft naar een maximale prestatie, terwijl een low-stakes afname conditie een leerling minder uitdaagt en eerder een typische prestatie zal uitlokken, in tegenstelling tot een maximale prestatie. In dit proefschrift bespreek we eerst verschillende data designs en de manier waarop ze normaal gesproken geïmplementeerd worden in de praktijk. Ook bespreken we de geschiktheid van de data designs om het effect van differentiële motivatie te kunnen onderzoeken. In Hoofdstuk 2 tot en met Hoofdstuk 5 onderzoeken we de geschiktheid van het mixture Rasch model en person-fit methoden om differentiële motivatie te modeleren. Het opleggen van specificaties aan de latente klassen in het mixture Rasch model zouden moeten helpen bij het identificeren van de latente klassen zodat één latente klas high-stakes antwoordgedrag vertegenwoordigt terwijl de ander low-stakes antwoordgedrag vertegenwoordigt.

In Hoofdstuk 2 gebruiken we data van een Nederlands toetsprogramma om te onderzoeken of de geschatte vaardigheidsverdelingen verkregen met twee operationele toetsen verschillen bij gebruik van verschillende data designs. De data designs verschilden met betrekking tot de afnameconditie van de ankeritems. We vonden dat het externe anker design robuuster is tegen het effect van differentiële motivatie dan het proeftoets design. Gebruik van het proeftoets design resulteerde in een substantiële overschatting van het verschil tussen de gemiddelden van de geschatte vaardigheidsverdelingen van de populaties die de operationele toetsen maakten. We controleerden voor het effect van differentiële motivatie in het proeftoets design door gebruik te maken van een mixture Rasch model om de operationele toetsen te linken. Het verwijderen van items die DIF vertonen tussen high-stakes en low-stakes afnamecondities resulteerde niet in een betere link.

Het doel van Hoofdstuk 3 was om te onderzoeken wanneer gebruik van het Rasch model als linkmethode in geval van differentiële motivatie tussen de stakes van een operationele toets en ankeritems een invalide link resultaat oplevert. Dit werd onderzocht voor een extern anker design en een variant van een proeftoets design. Ook werd onderzocht of een mixture Rasch model met specificaties voor de latente klassen de gesimuleerde latente klassen kan identificeren. De resultaten gaven aan dat voor een extern anker design het link resultaat verkregen met het Rasch model bias vertoont wanneer de hoeveelheid gemotiveerde leerlingen verschilt tussen de subpopulaties waarbij de anker items worden afgenomen. Echter, het mixture Rasch model met specificaties voor de latente klassen identificeerde niet de gesimuleerde latente klassen die low-stakes en high-stakes antwoordgedrag moeten vertegenwoordigen. Wanneer een proeftoets design gebruikt werd om de operationele toetsen te linken, vonden we bias in alle condities. De hoeveelheid bias nam toe met het percentage leerlingen die low-stakes antwoordgedrag vertoonden. Het mixture Rasch model met specificaties voor de latente klassen identificeerde alleen in een beperkt aantal condities de gesimuleerde latente klassen die low-stakes en high-stakes antwoordgedrag vertegenwoordigen.

In Hoofdstuk 4 bespreken we in hoeverre een mixture Rasch model met specificaties voor de latente klassen en de l_z person-fit methode gebruikt kunnen worden om verschillen in motivatie in een low-stakes afnameconditie te modeleren. We onderzochten de geschiktheid van een mixture modeling strategie in een steekproef van 1.512 basisschoolleerlingen door de posteriori kansen van het mixture model met specificaties op de latente klassen te vergelijken met het door leerlingen zelf gerapporteerde motivatieniveau. Ook onderzochten we de relatie tussen het door leerlingen zelf gerapporteerde motivatieniveau en de l_z person-fit methode. Op basis van de resultaten concludeerden we dat, vergeleken met de posteriori kansen van het mixture Rasch model met specificaties op de latente klassen, het gebruik van de l_z person-fit methode een betere methode is voor het modeleren van verschillen in motivatie.

Empirisch onderzoek geeft aan dat het ontbreken van motivatie tijdens een toets afname een negatief effect kan hebben op de toetsscores van leerlingen en op de consistentie van het antwoordpatroon. In Hoofdstuk 5 werd het verschil onderzocht tussen antwoordgedrag in low-stakes en high-stakes afnamecondities met betrekking tot toetsprestaties en consistentie van het antwoordpatroon. Een verschil tussen consistentie van het antwoordpatroon werd vaker gevonden dan een verschil tussen toetsprestaties. Zelden werd gevonden dat leerlingen verschillen op consistentie van het antwoordpatroon en toetsprestaties. Toetsmotivatie gemeten met een vragenlijst is een significante voorspeller van variatie in (1) consistentie in antwoordgedrag op een toets afgenomen in een low-stakes conditie en (2) het verschil in prestatie tussen de low-stakes en de high-stakes toets. Echter, de hoeveelheid verklaarde variantie is gering.

In Hoofdstuk 6 lichten we toe waarom het gebruik van (1) een meer omvattende operationalisering van differentiële motivatie en (2) andere statistische modellen bruikbaar kan zijn in verder onderzoek naar differentiële motivatie en het modeleren ervan.

Woord van dank

Hierbij wil ik alle mensen bedanken die hebben bijgedragen aan de totstandkoming van dit proefschrift. In het bijzonder wil ik mijn promotor en copromotor bedanken. Klaas, ik ben erg blij dat we onze samenwerking na mijn bachelor's thesis hebben voortgezet en dat het heeft geleid tot dit proefschrift. Bedankt voor je vertrouwen in mij. Anton, jij bood mij de kans om dit promotieproject bij Cito uit te voeren. Heel erg bedankt voor je steun.

Verder wil ik mijn collega's bij POK en het MTO-departement bedanken. In het bijzonder wil ik Maaike, Anke, Ron, Saskia, Renske, Zsuzsa en Pieter bedanken voor de ontspannende koffie-momentjes, mooie tijden op conferenties en natuurlijk voor alle vormen van input voor mijn proefschrift. Wilco, naast alle gezellige momenten wil ik je bedanken voor de fijne samenwerking aan het person-fit paper. Margot en Hendrik, bedankt voor alle kleine momenten van feedback en tips en vooral voor het leuker maken van ons promotietijdperk!

Ook wil ik mijn familie bedanken. Mam, dankzij jou kon ik de afgelopen maanden meer tijd besteden aan mijn manuscript. Bedankt voor alle steun. Cynthia, heel erg bedankt voor het ontwerpen van de omslag van dit proefschrift. Joshua, door jou valt alles te relativeren. Wat een prachtig geschenk is dat! En tot slot, Richard, de afgelopen vier jaar is er bijzonder veel gebeurd. Ik zou niet weten hoe ik het zonder jou had moeten doen. Simpelweg bedankt voor alles.