

## **Comparing the Effectiveness of Different Linking Designs: the Internal Anchor versus the External Anchor and Pre-Test Data**

**Marie-Anne Mittelhaeuser  
Anton A. Béguin  
Klaas Sijtsma**





**Comparing the Effectiveness of Different Linking Designs:  
the Internal Anchor versus the External Anchor and  
Pre-Test Data**

Marie-Anne Mittelhaäuser, Tilburg University & Cito

Anton A. Béguin, Cito

Klaas Sijtsma, Tilburg University

Cito  
Arnhem, 2011

This manuscript has been submitted for publication. No part of this manuscript may be copied or reproduced without permission.

## Abstract

The goal of the current study was to compare a linking procedure for two test forms using different types of common items. It was hypothesized that the test-taking condition of the common items influences the linking procedure. The results support the hypothesis. A mixed Rasch model was used to model some examinees as being more motivated than others to solve the items. Removal of aberrant item-score vectors or items displaying differential item functioning did not improve the linking procedure.

## Comparing the Effectiveness of Different Linking Designs: the Internal Anchor versus the External Anchor and Pre-Test Data

Test forms may differ with respect to the difficulty of the test and, as a result, scores on different test forms are often not directly comparable. Several procedures such as linking, scaling and equating are available to develop a common metric (see e.g. Kolen & Brennan, 2004). Different linking procedures are based on different assumptions and use different designs or data collection procedures. One design particularly useful in educational testing is the common-item non-equivalent groups design. In this design, two high-stakes test forms are administered in two different populations, for example, eighth-grade primary-school students in two successive years, and both test forms are linked by means of common items. The test forms to be linked by means of the common items are also called operational test forms. In an educational testing context, populations often are not equivalent and the proficiency level of examinees may change from year to year. The common-item non-equivalent groups design can accommodate these problems to produce a common scale. A popular choice for the common items is to let them represent a miniature version of the total test form. In an item response theory (IRT) context, this means that the items measure the same latent proficiency and that the same IRT model has to fit the common items and the total test form.

In this study, three types of common-item non-equivalent groups designs are discussed. The difference between the designs most relevant to this study concerns the issue whether common items are used for determining the total test score and thus whether they are administered under high-stakes conditions. Figure 1 shows symbolic representations of the designs. In this figure, rows correspond to examinee data and columns to item data. Boxes represent combinations of items and examinees for which data are available. The order of the items presented in the figures might not correspond to the order in the real test.

Figure 1a shows the internal anchor design. In this design, samples of both populations of examinees are administered a different test form, which in both cases includes a selection of items from the test form to be linked, also referred to as operational items, and additional common items, in this design also referred to as internal anchor items. The common items are the same across the different test forms and the different samples of examinees. Therefore, differences in difficulty between test forms can be estimated based on the relative

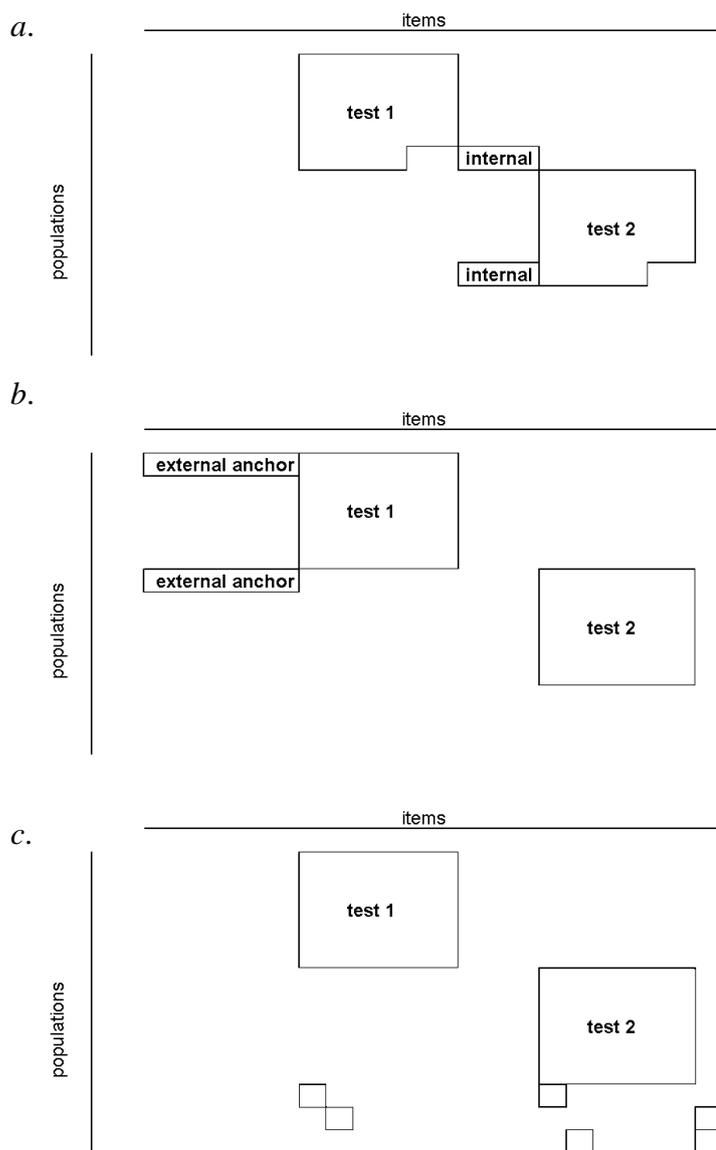


Figure 1. (a) The internal anchor design, (b) the external anchor design and (c) the pre-test design.

performance of a sample on the common items. Both tests can be linked to the common items, and therefore indirectly to each other. In this design, all items including the common items contribute to an examinee's total test score. Therefore, the administration conditions for both the items in the operational test and the common items are high-stakes. Test security might be threatened if every examinee in both tests is presented the common items. However, in the internal anchor design, only samples of both populations are presented the common items and the examinees in this sample are not aware that they are presented an alternative test form. Furthermore, the internal anchor items, the number of internal anchor items and the number of examinees who are presented the internal anchor items are not made public. It can thus be assumed that the threat to test security is minimal when using the internal anchor design. Therefore, the internal anchor design is especially useful in linking two high-stakes test forms. Placing the internal anchor items in the same position in the total test in both operational test forms can avoid undesirable order effects when the different operational test forms are compared.

The external anchor design (Figure 1b) is different from the internal anchor design in that the score on the common items does not contribute to the total test score and that the external anchor does not replace part of the test. Therefore, when using an external anchor design to link two test forms, the link between the tests is based on the additional common items, in this design also referred to as external anchor items. Usually, examinees know that the external anchor items do not contribute to their total test score and, as a result, these items are administered in a low-stakes condition whereas the operational items are administered in a high-stakes condition. Proficiency differences between external anchor items and operational items may be attributed to an administration effect.

The third method is the pre-test design (Figure 1c). Subsets of items intended for use in the operational tests may be pre-tested on different samples of examinees so as to examine

the statistical characteristics of items before including them in an operational test. Items with the most promising item characteristics are selected into the operational test. However, if items are pre-tested in two consecutive years, they can be used as an external link between two operational tests. Hence, similar to the anchor in the external anchor design, the pre-test is administered in a low-stakes condition.

Klein and Jarjoura (1985) concluded that it was important to use content-representative common items. Therefore, a popular choice for the items is to let them represent a miniature version of the total test form. In an IRT context, this means that the items measure the same latent proficiency and that the same IRT model fits the common items and the total test form. However, Wise and DeMars (2005) found that if item performance does not contribute to the total score, examinees might not give their best effort. Consequently, the performance on the common items may differ from performance on the total test form due to different test taking conditions. This may result in unusual patterns of item scores or in relatively meagre performance, and the effect may be an increase in uncertainty in the linking procedure using an external anchor or pre-test items compared to using internal anchor items.

A mixed IRT model may be used to test if some of the item-score vectors (i.e., the vector of item scores an examinee has produced) have been affected differently by the low-stakes condition compared to the high-stakes operational test. Mixed 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. 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 different test forms, one can specify the mixed IRT model in such a way that one of the latent classes represents high-stakes response

behavior (represented by vectors of item scores unique to this latent class) while the other latent class represents low-stakes behavior (Béguin, 2005; Béguin & Maan, 2007). If only the data of the high-stakes class are used in the linking procedure, this is expected to improve the results of the linking procedure.

Let  $X_i$  denote the score on item  $i$ , with the total number of items represented by  $k$ .

According to the mixed IRT model, the probability of passing item  $i$  ( $X_i = 1$ ) depends on a class-specific person parameter  $\theta_{jg}$ , denoting the proficiency of examinee  $j$  if he/she belongs to latent class  $g$ . The techniques currently available for estimating a mixed IRT model focus on the Rasch model. The limitation to the Rasch model is partly due to the limited information in the data to estimate more-complex models. The mixed 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  $\beta_{ig}$  is a class-specific difficulty parameter. Aggregated over items, the probability of obtaining an item-score vector  $\mathbf{x}_j = \{x_{1j}, x_{2j}, \dots, x_{kj}\}$  given proficiency  $\theta_j$  and membership of class  $g$  is

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

Let  $\pi_g$  denote the proportion of the population that belongs to class  $g$  ( $g = 1, \dots, G$ ).  $\pi_g$  is also called the class probability. The probability for an individual  $j$  to belong to class  $g$  depends on the item-score vector in the following way:

$$p(g | \mathbf{x}_j) = \frac{\pi_g p(\mathbf{x}_j | g)}{\sum_{g=1}^G \pi_g p(\mathbf{x}_j | g)}. \quad (1)$$

Instead of removing the latent class displaying low-stakes response behavior from the linking procedure, an alternative procedure to improve the link is to remove item-score vectors from the data, to which an IRT model does not fit. This can be done using a person-fit statistic, which attempts to assess the fit of the IRT model at the individual level (Embretson & Reise, 2000). A person-fit statistic developed to assess the likelihood of an item-score vector is the  $l_z$  statistic (Drasgow, Levine & Williams, 1985; Meijer, 2003). The  $l_z$  statistic is given by

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

where  $l$  denotes the unstandardized likelihood of the item-score vector, and  $E(l)$  and  $\text{Var}(l)$  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

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

The  $l_z$  statistic is often assumed to be a standard normal deviate, with large negative values indicating misfit. Van Krimpen-Stoop and Meijer (1999) showed that the normal approximation to  $l_z$  is invalid, yielding a conservative test, in particular for detecting aberrant item-score vectors at the lower and higher ends of the scale and when applied to short scales. Fortunately, Snijders (2001) derived an asymptotic null distribution where the proficiency parameter is replaced by an estimate.

We used data from a Dutch testing program to investigate if the link between two operational tests differs when using different types of common items. This was done by comparing the mean proficiency differences of the operational tests over the different types of common items. The differences in mean proficiency of the operational tests were assessed by means of confidence intervals constructed using a bootstrap procedure. Furthermore, it was evaluated if item-score vectors, which were influenced most by the low-stakes test-taking conditions, could be identified. For this purpose, we used both the mixed Rasch model and the  $l_z$  person-fit statistic. Subsequently, Item-score vectors identified using the mixed Rasch model or the  $l_z$  statistic were removed from the data to improve the link between the two operational tests forms. Furthermore, item-misfit was investigated in order to inspect if removing misfitting items would improve the link between the two operational test forms.

## Method

### *Participants and Design*

Data were used from the reading and mathematics scales of the ‘*Eindtoets Basisonderwijs*’ (End of Primary Education Test). This test is administered every year at the end of Dutch primary education, and the examinees’ results are used to give advice about the most appropriate type of secondary education. Test administration is high stakes and item secrecy is vital; hence, the test form is renewed each year. A link between different test forms can be established using an internal anchor, an external anchor and pre-test data. We developed a common metric for the mathematics and reading scales using two consecutive test forms, which are the operational test forms administered in 2009 and 2010.

Both operational test forms contained 30 reading items and 60 mathematics items. Samples contained 4,995 participants for 2009 and 5,123 participants for 2010. The internal anchor test consisted of 14 reading items and 20 mathematics items, which were administered to 2,989 and 2,421 participants in 2009 and 2010, respectively. The external anchor test

consisted of 15 reading items and 20 mathematics items, which were administered to 5,086 and 4,575 participants in 2009 and 2010, respectively. In order to pre-test the reading items for the 2009 operational test, in 2008 16 reading pre-test booklets (ranging from 29 to 31 items) and 19 mathematics pre-test booklets (ranging from 30 to 90 items) were administered. The number of participants who were administered the reading pre-test booklets ranged from 185 to 310, and for the mathematics items 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 and ranged from 440 to 607 for the reading items and from 219 to 1,685 for the mathematics items. The items for the 2010 operational test were pre-tested in 2009 using 17 reading pre-test booklets containing reading items (ranging from 29 to 32 items) and 23 mathematics pre-test booklets (ranging from 29 to 60 items). The number of participants who were administered these pre-test booklets ranges from 220 to 373 for the reading items and from 46 to 372 for the mathematics items. The number of observations per item ranged from 457 to 995 for the reading items and 504 to 1,664 for the mathematics items.

### *Analyses*

To inspect differences in mean proficiency, the Rasch Model was fitted to the data (Rasch, 1960). According to the Rasch model, the probability of passing item  $i$  for individual  $j$  is a function of proficiency parameter  $\theta_j$  and can be given by

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

where  $\beta_i$  is the difficulty parameter of item  $i$ . The fit of the Rasch model to the operational test items was investigated by means of Infit and Outfit statistics (Wright & Masters, 1982) available in the eRm package in R (Mair, Hatzinger & Maier, 2010). Items having a Mean Square Outfit statistic or Mean Square Infit statistic outside the range of 0.5 – 1.5 were iteratively removed from the analyses. The OPLM software was used to estimate the Rasch

model. The differences in mean proficiency of both operational tests were compared for each of the three linking designs. Student's *t*-tests were used to determine whether the differences between the means of the operational tests of 2009 and 2010 were significant. Cohen's *d* was used to assess effect size (Cohen, 1988).

The standard deviations of the proficiency distributions provided by OPLM may be used to evaluate the significance of the differences between the mean proficiencies of the operational tests of 2009 and 2010. However, instead of using the complete variance-covariance matrix of the difficulty parameters, 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. Therefore, a bootstrap procedure (Efron & Tibshirani, 1993) was used to construct 95% confidence intervals for the differences between the mean proficiencies of the operational tests. The bootstrap procedure was done using the following steps:

1. For the internal anchor, external anchor and pre-test data, 1000 bootstrap samples were drawn. The operational test data matrix was kept constant. Bootstrapping was done using the statistical program R (R Development Core Team, 2005).
2. OPLM was used to estimate the mean proficiency of the examinees for each operational test. A batch file was used to repeat this sequence of analyses for each bootstrap sample.
3. Steps 1 and 2 resulted in 1000 differences in mean proficiency between operational tests for each type of linking data. The Shapiro-Wilk test (Shapiro & Wilk, 1965) was used to test whether differences found for each type of linking data were normally distributed. A 95% confidence interval (CI) was constructed using the .025 and .975 percentiles under the normal distribution.

Internal anchor items are administered in a high-stakes condition, and only item-score vectors of the external anchor items and pre-test data that were both administered in a low-stakes condition, are expected to belong to either a latent class displaying low-stakes response behavior or a latent class displaying high-stakes response behavior. Therefore, the mixed Rasch model was only estimated for the external anchor design and the pre-test design. A dedicated version of the OPLM software (Béguin, 2008) estimated the mixed Rasch model. Two latent classes were specified in the mixed IRT model, the first representing response behavior expected under high-stakes conditions and the second allowing for response behavior typical of less motivated examinees. The item-score vectors of the operational tests were modelled as being exclusively part of the first latent class, which was done by setting  $\pi_0 = 0$  and  $\pi_1 = 1$  in Equation 1 for all item-score vectors of the operational tests. The item-score vectors of the external anchor data and the pre-test data could be either in the first or the second latent class. The mixed Rasch model was compared to the simple Rasch model to investigate if modelling subpopulations would improve the link between the operational tests. This was done by means of (1) comparing difficulty parameters estimated for both the Rasch model and the mixed Rasch model, (2) a test for model comparison, and (3) by comparing the differences in mean proficiency between the samples administered the two operational tests.

Next to the application of mixed IRT models, a data cleaning procedure was used to investigate how removing aberrant item-score vectors influences the external anchor link of the operational tests. These item-score vectors were removed as follows:

1. First, the operational-test items were used to estimate the proficiency parameter  $\theta$  for each examinee.
2. Second,  $\theta$  estimated for the operational items was used to compute the  $l_z$  statistic on the external anchor items for each examinee. The asymptotic null distribution developed by Snijders (2001) was used to correct for the

conservative nature of the  $l_z$  statistic . This was done using program R (R Development Core Team, 2005).

3. Finally, the item-score vectors corresponding to the lowest 1%, 5%, 10%, 25% and 50%  $l_z$  statistics were removed from the data.

After removing the aberrant item-score vectors, the Rasch model was fitted to the data and the mean proficiency differences of the samples administered the operational tests were compared.

Instead of removing examinees whose item-score vectors were affected by administering low-stakes common items, one could also remove items that function differently in different groups of examinees. Differential Item Functioning (DIF) identifies items that display different statistical properties in different group settings after controlling for differences between the proficiencies of the groups (Holland & Wainer, 1993). Items displaying DIF between the high-stakes condition and the low-stakes condition are not suited for establishing a common metric between the two operational test forms. The OPLM software provides the contribution of each item to the  $R_{1c}$  statistic (Glas, 1989), which evaluates the difference between expected and observed proportions of item scores in homogeneous score groups. Items having a mean sum of squares in excess of 4 were selected for visual inspection of DIF. OPLM provided graphs displaying the Item Characteristic Curve (ICC) for different groups. After removing items displaying DIF, the Rasch model was again fitted to the data and the differences between mean proficiencies of the operational tests were compared when linked each type of common items.

## Results

### *Rasch Analysis*

One reading item of the operational test of 2009 was deleted from further analyses because the Outfit Mean Square value was 1.519. Furthermore, one pre-test reading item was

removed from the analyses after inspection of the ICC because of extreme misfit. Table 1 shows the estimated latent proficiency means of the operational tests for reading and of the operational tests for mathematics. The mean of the 2010 operational reading test was higher than the mean of the 2009 operational reading test for each type of common items ( $p < .01$ ). The mean of the 2010 operational mathematics test was significantly higher than the mean of the 2009 operational mathematics test for the internal anchor design ( $p < .05$ ), the external anchor design ( $p < .01$ ) and the pre-test design ( $p < .01$ ). However, the effect sizes of the external anchor and pre-test design were considerably higher than those of the internal anchor design. Therefore, further inspection of the use of different types of linking data is desirable.

#### *Confidence Intervals*

A bootstrap procedure was used to construct 95% confidence intervals for the differences between mean proficiencies of the operational test of 2009 and 2010; see Table 1, Column 6. The Shapiro-Wilk test (Shapiro & Wilk, 1965) indicated that the differences found for each type of linking data were normally distributed ( $p > .05$ ). The mean proficiencies of the operational reading tests differed significantly with each type of linking data, which is consistent with the Rasch analysis. The confidence intervals constructed for the mathematics items differ for the different linking designs. For example, the confidence interval for the internal anchor design (0.014; 0.090) does not overlap with the confidence interval for the external anchor design (0.093; 0.156) or the pre-test design (0.091; 0.160). Even though the confidence intervals constructed for the reading items do overlap, the boundaries and widths of the confidence intervals for the different linking designs differ. Therefore, we have reason to conclude that the results of the linking procedure for the mathematics tests and the reading tests differ between the types of common items used.

#### *Mixed IRT*

The difficulty parameters of both latent classes obtained under the mixed Rasch model

Table 1.

*Proficiency Distributions of Operational Tests Using Different Types of Linking Data*

Linking data	Population	<i>M</i>	<i>SD</i>	Cohen's <i>d</i> / Sign. Student's <i>t</i>	95% CI
Reading items					
Internal anchor	2009	1.453	0.832	0.124 / **	(0.057;0.165)
	2010	1.565	0.967		
External anchor	2009	1.610	0.825	0.194 / **	(0.132;0.217)
	2010	1.784	0.965		
Pre-test	2009	1.539	0.824	0.155 / **	(0.095;0.204)
	2010	1.677	0.955		
Mathematics items					
Internal anchor	2009	1.169	1.023	0.049 / *	(0.014; 0.090)
	2010	1.220	1.071		
External anchor	2009	1.118	1.021	0.120 / **	(0.093; 0.156)
	2010	1.243	1.069		
Pre-test	2009	1.065	1.021	0.120 / **	(0.091; 0.160)
	2010	1.190	1.064		

\*  $p < .05$ . \*\*  $p < .01$ .

were compared with the difficulty parameters of the Rasch model. The difficulty parameters of both models are comparable because the mean of the proficiency distribution of the 2010 operational test is fixed at 0 in both models. Figure 2 shows for a pre-test design that the difficulty parameter estimates of the reading items in the motivated class were lower than in the unmotivated class. The difficulty parameters estimated in the motivated class were

approximately the same when estimated by the mixed Rasch model and the simple Rasch model. However, four item difficulties from the mixed Rasch model were lower than  $-7$ , which deviated from the general trend. These items were identified as items that were answered correctly by almost every examinee in the motivated condition. Figure 3 shows that the item difficulties were higher for the unmotivated class than the motivated class when the pre-test design was used to link the operational mathematics tests. This suggests that examinees in the unmotivated class find it more difficult to answer an item correct than examinees in the motivated class. Figure 4 (reading) and Figure 5 (mathematics) show that the item difficulties in the motivated class are smaller than in the unmotivated class when using an external anchor design.

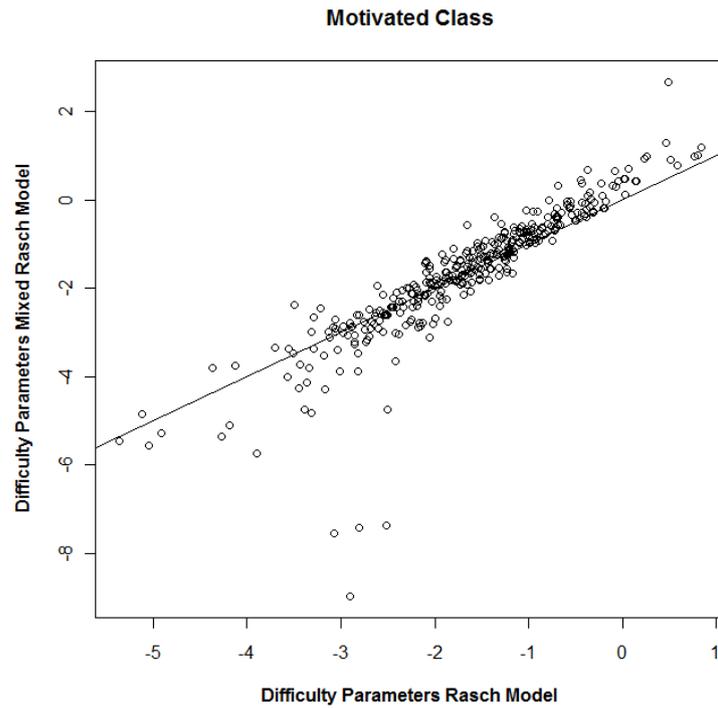
The class-membership probabilities for the motivated class were estimated for the external anchor tests and the pre-tests. For the external anchor tests, the probabilities were approximately equal in 2009 and 2010. For example, for the external-anchor reading test the probability for 2009 was .78 and for 2010 it was .77. For the mathematics tests the probabilities were .62 (2009) and .63 (2010). The class-membership probabilities for the motivated class differed among the pre-test booklets. The mean probability of the 14 reading pre-test booklets<sup>1</sup> to pre-test the items of 2009 was .58 (ranging from .18 to .83) whereas the mean probability of the 17 reading pre-test booklets to pre-test the items of 2010 was .58 (ranging from .45 to .74). The mean probability of the 18 mathematics pre-test booklets<sup>2</sup> to pre-test the items of 2009 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).

---

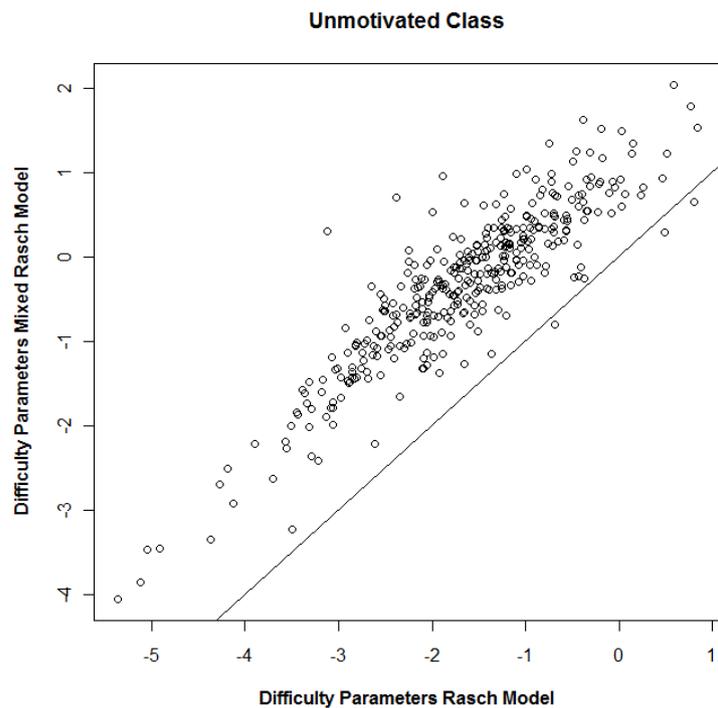
<sup>1</sup> Two pre-test booklets were iteratively removed from the analysis because the probability to belong to the motivated class was .00000 and .00006

<sup>2</sup> One pre-test booklet was removed from the analysis because the probability to belong to the motivated class was .00000

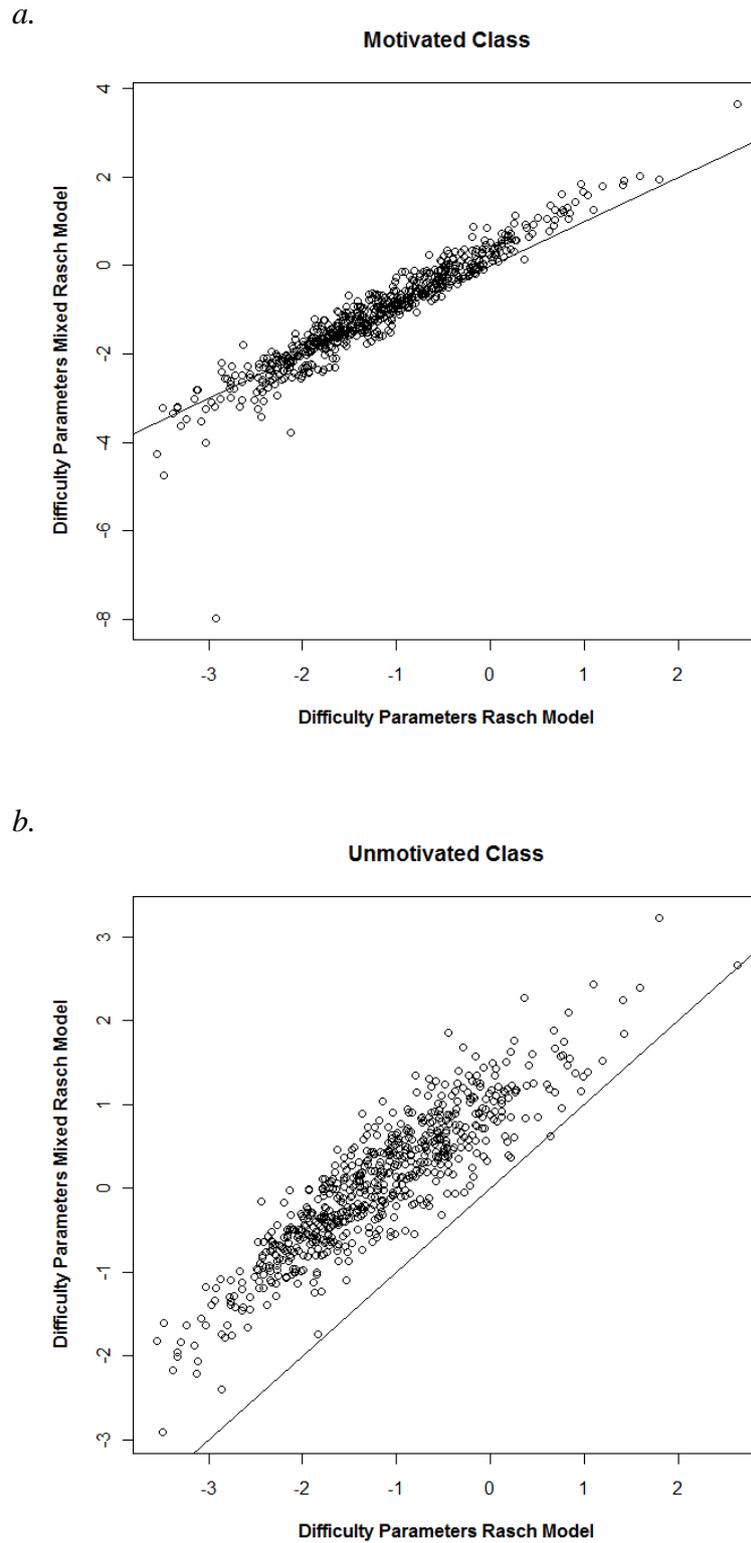
a.



b.

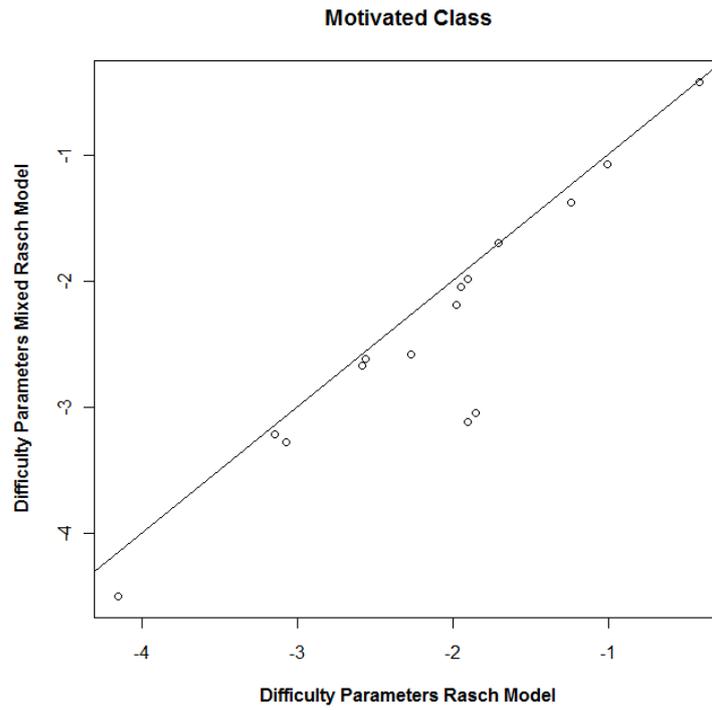


*Figure 2.* Item difficulty parameters of the Rasch model and the mixed Rasch model for the motivated class (a) and the unmotivated class (b) estimated in the pre-test design for the reading items.

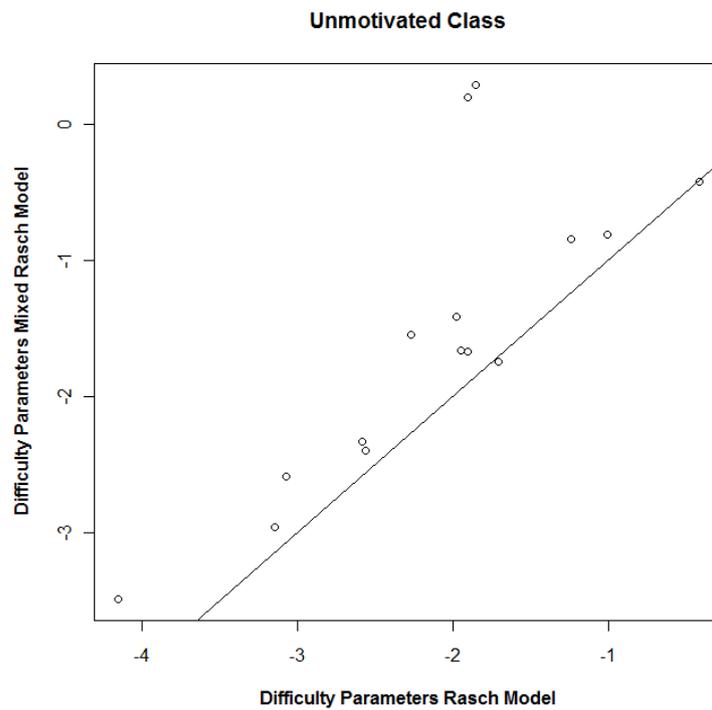


*Figure 3.* Item difficulty parameters of the Rasch model and the mixed Rasch model for the motivated class (a) and the unmotivated class (b) estimated in the pre-test design for the mathematics items.

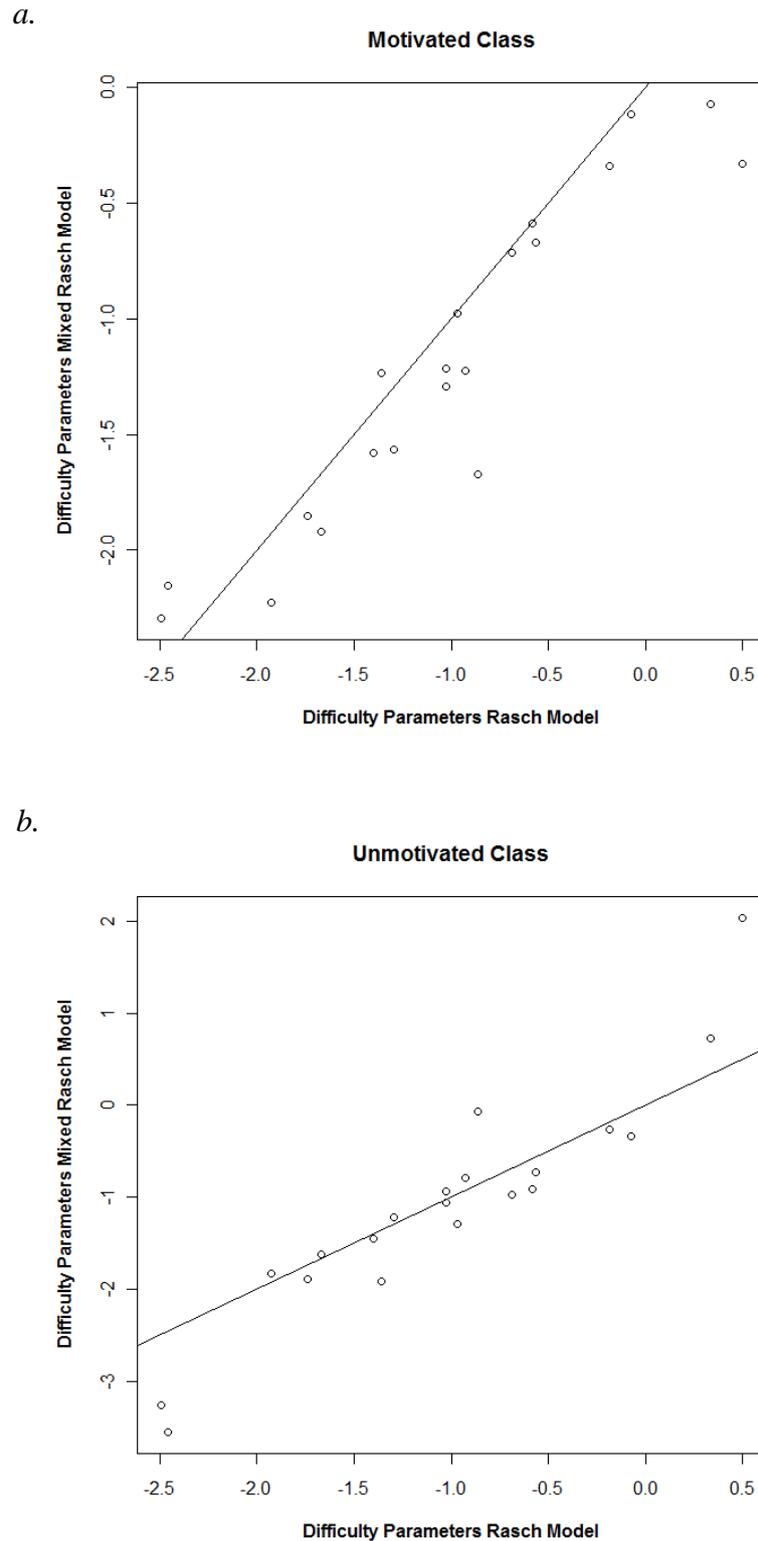
a.



b.



*Figure 4.* Item difficulty parameters of the Rasch model and the mixed Rasch model for the motivated class (a) and the unmotivated class (b) estimated in the external anchor design for the reading items.



*Figure 5.* Item difficulty parameters of the Rasch model and the mixed Rasch model for the motivated class (a) and the unmotivated class (b) estimated in the external anchor design for the mathematics items.

Based on the log likelihoods, the mixed Rasch model provided a better fit to the data than the simple Rasch model. This was found when the pre-test design was used to link the reading tests ( $\chi^2(368, N = 19,111) = 10,714.9, p < .001$ ), and the mathematics test ( $\chi^2(647, N = 21,753) = 42,977.7, p < .001$ ), and when the external anchor design was used to link the reading tests ( $\chi^2(73, N = 19,779) = 7,270.8, p < .001$ ) and the mathematics tests ( $\chi^2(138, N = 19,779) = 5,799.2, p < .001$ ).

Since the results of the mixed IRT analysis indicated that the mixed Rasch model with two latent classes provided a better fit to the data than the simple Rasch model, we decided to investigate the effect of assuming two subpopulations to be present in the external anchor data and pre-test data on the mean proficiency of both operational tests. Table 2 shows that for the reading tests the differences between mean proficiencies of the operational tests increased slightly when the external anchor design was used. The differences between mean proficiencies of the mathematics tests linked with the external anchor items remained the same. However, the differences between mean proficiencies decreased compared to the simple Rasch model when the pre-test design was used. This was found for the reading tests (Cohen's  $d$  was smaller by .123) and the mathematics tests (Cohen's  $d$  was smaller by .119).

#### *Person-misfit*

Five different datasets were constructed by consecutively removing examinees with the 1% lowest  $l_z$  statistics (first dataset), the 5% lowest  $l_z$  statistics (second dataset), and then the 10%, 25% and 50% lowest  $l_z$  statistics (third, fourth and fifth datasets). The Rasch model was estimated in each dataset to inspect how removal of aberrant item-score vectors affected the differences in mean proficiency between the operational tests. Table 3 shows that the difference in mean proficiency between the operational reading tests increased slightly compared to not removing aberrant item-score vectors. Removal of aberrant item-score

Table 2.

*Proficiency Distributions of Operational Tests Estimated with the Mixed Rasch Model.*

Model	Linking data	Population	$M$	$SD$	Cohen's $d$ / Sign. Student's $t$
Reading items					
Mixed Rasch	External Anchor	2009	-0.183	0.823	0.204 / **
		2010	0.000	0.965	
	Pre-test	2009	0.029	0.829	0.032
		2010	0.000	0.967	
Mathematics items					
Mixed Rasch	External Anchor	2009	-0.125	1.021	0.120 / **
		2010	0.000	1.069	
	Pre-test	2009	0.008	0.885	0.001
		2010	0.000	1.070	

\*  $p < .05$ . \*\*  $p < .01$ .

vectors did not seem to have an effect on the proficiency differences of the operational mathematics tests.

*Differential Item Functioning*

Inspection of the  $R_{1c}$  and ICCs suggested that the internal and external anchor items for both the reading test and mathematics test did not show DIF. Visual inspection of the ICCs suggested that higher values of the  $R_{1c}$  statistic was not due to DIF, but rather to a general lack of model fit. However, linking the two operational reading tests with pre-test items resulted in six items displaying DIF. Furthermore, linking the two operational mathematics tests resulted in eight items displaying DIF.

Table 3.

*Proficiency Distributions of Operational Tests without Aberrant Item-Score Vectors.*

Percentage	Population	<i>M</i>	<i>SD</i>	Cohen's <i>d</i> / Sign. Student's <i>t</i>
Reading items				
1	2009	1.613	0.826	0.198 / **
	2010	1.791	0.965	
5	2009	1.627	0.826	0.201 / **
	2010	1.808	0.966	
10	2009	1.641	0.827	0.211 / **
	2010	1.831	0.966	
25	2009	1.694	0.828	0.202 / **
	2010	1.876	0.968	
50	2009	1.806	0.831	0.202 / **
	2010	1.989	0.971	
Mathematics items				
1	2009	1.121	1.021	0.118 / **
	2010	1.244	1.069	
5	2009	1.126	1.022	0.119 / **
	2010	1.250	1.070	
10	2009	1.135	1.022	0.113 / **
	2010	1.253	1.070	
25	2009	1.141	1.023	0.129 / **
	2010	1.276	1.071	
50	2009	1.171	1.025	0.122 / **
	2010	1.299	1.072	

\*  $p < .05$ . \*\*  $p < .01$ .

After removal of the reading items and mathematics items displaying DIF in the pre-test data, the OPLM was fitted to the remaining data again. Table 4 shows the mean proficiencies of the operational tests estimated without DIF items. Removal of items displaying DIF did not seem to have an effect on the proficiency differences of the operational reading test and the operational mathematics test.

### Discussion

We conclude that the results of the linking procedure depend on the type of common items. For example, the confidence interval for the mean proficiency difference of the operational mathematics tests constructed with internal anchor items did not overlap with the confidence intervals constructed with the external anchor items and pre-test items. We also found evidence for the existence of differently motivated subpopulations. Removal of misfitting item-score vectors and items hardly affected the linking procedure. The conclusions were roughly the same for reading items and mathematics items.

As a result of fitting the mixed Rasch model, we found for an external anchor that the differences in mean proficiency between the two operational tests did not change but that the differences seemed to disappear for the pre-test data. Class-membership probabilities of the different tests for the motivated class might explain these results. The class-membership probabilities were almost the same in both years, both for the reading items and the mathematics items. However, the class-membership probabilities of the pre-test booklets varied tremendously, which made it worthwhile to fit a mixed Rasch model.

The current study used data from a test for which both internal and external anchor items as well as pre-test data are available. One could ask whether a combination of types of common items provides the best link. This is an interesting question to investigate in future research, but for all tests with a limited linking design, it is interesting to know what the effectiveness of the particular types of linking designs is.

Table 4.

*Proficiency Distributions of Operational Tests Using Different Types of Linking Data Without DIF Items*

Linking data	Population	<i>M</i>	<i>SD</i>	Cohen's <i>d</i> / Sign. Student's <i>t</i>
Reading items				
Internal anchor	2009	1.453	0.832	0.124 / **
	2010	1.565	0.967	
External anchor	2009	1.610	0.825	0.194 / **
	2010	1.784	0.965	
Pre-test	2009	1.559	0.836	0.186 / **
	2010	1.728	0.975	
Mathematics items				
Internal anchor	2009	1.169	1.023	0.049 / **
	2010	1.220	1.071	
External anchor	2009	1.118	1.021	0.120 / **
	2010	1.243	1.069	
Pre-test	2009	1.066	1.031	0.115 / **
	2010	1.187	1.081	

\*  $p < .05$ . \*\*  $p < .01$ .

The use of a mixed Rasch model proved to be useful with the current data. However, since the assumption of equal discrimination for all items is only partially valid and not likely to be met in most real datasets, it might be interesting to develop a mixed Birnbaum model.

Furthermore, the current research only investigated a mixed 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 multidimensional IRT model is more appropriate to model response behavior (Embretson & Reise, 2000).

## References

- 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., & 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)
- 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)
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (second ed.)  
Lawrence Erlbaum Associates.
- Dragow, 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.
- 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.
- Embretson, S. E., & Reise, S. P. (2000). Assessing the fit of IRT models. In *Item response theory for psychologists* (pp. 226-246). Mahwah, NJ: Lawrence Erlbaum.
- Glas, C. A. W. (1989). *Estimating and testing Rasch models*. Doctoral dissertation, University of Twente.
- Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum.
- 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*. (2<sup>nd</sup> ed.). New York, NY: Springer Verlag.
- Mair, P., Hatzinger, R., & Maier, M. (2010). *eRm: Extended Rasch Modeling*.  
<http://CRAN.R-project.org/package=eRm>.
- Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods*, 8, 72-87.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*.  
Copenhagen: Danish Institute for Educational Research.
- R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing*. [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.
- Rost, J. (1997). Logistic Mixture Models. In: W. van der Linden & R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 449-463).
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591-611.
- Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika*, 66, 331-342.
- 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.
- 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, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: an extension of the generalized partial-credit model. *Applied Psychological Measurement*, 28, 389-406.

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

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