MEASURING AND DETECTING DIFFERENTIAL ITEM FUNCTIONING IN CRITERION-REFERENCED LICENSING TEST

A Theoretic Comparison of Methods

Marie Wiberg

EM No 60, 2007

UMEÅ UNIVERSITET

ISSN 1103-2685

Abstract

The validity of a measurement instrument depends on the quality of the items included in the instrument. The overall aim was to compare methods for detecting and measuring differential item functioning, DIF, in order to find a suitable method for examining DIF in a dichotomously scored criterion-referenced licensing test. The methods were discussed with respect to whether they are parametric, the nature of the matching score, if they can handle dichotomously and polytomously scored items, if they can test and/or measure DIF, and if they can detect both uniform and non-uniform DIF. The methods were also discussed with respect to whether they could handle the cut-off score in particular and the sample size requirements. The results show that there is not one method that can be recommended because many of them rely on strong assumptions which need to be examined and fulfilled before they can be recommended. It was recommended that an empirical study comparing the Mantel-Haenszel, logistic regression, log linear models and an IRT method is performed. Finally, the concluding remarks provide a discussion of guidelines for what to do if an item displays DIF in a test.

Table of content

1. Introduction

In order to draw valid conclusions from an achievement test it is essential that the test is a valid measurement of what it is intended to measure. See e.g. Messick (1989) or Kane (2006) for a discussion of validity in tests and validation, i.e. the process of ensuring the validity in a test with respect to the context it is used. A test is never better than the sum of its items; hence to identify problematic items through item analysis is of great importance. Item analysis includes using statistical techniques to examine the test takers' performance on the items. One important part of the item analysis is to examine Differential Item Functioning, DIF, in the items. There exists several definitions of DIF but the following will be used in this report:

> "DIF refers to differences in item functioning after groups have been matched with respect to ability or attribute that the item purportedly measures… DIF is an unexpected difference among groups of examinees who are supposed to be comparable with respect to attribute measured by the item and the test on which it appears" (p.37) (Dorans & Holland, 1993).

Angoff (1993) points out that an item which displays DIF has different statistical properties in different group settings when controlling for differences in the abilities of the groups. It is therefore important to use representative samples in order to draw valid conclusions about DIF. Detection of DIF in items in a test is important regarding the quality of an assessment instrument since DIF can be described as the presence of nuisance dimensions intruding on the ability intended to measure (Ackerman, 1992). It is important to stress that DIF is an *unexpected* difference between two groups after matching on the underlying ability that the item is intended to measure. Note that DIF is synonymous with statistical bias, i.e. the under- or over-estimation of one or more parameters in the statistical model (Camilli, 2006).

Before proceeding, it is important to clarify one concept which has been used previously instead of DIF but now has another meaning; item bias (Scheuneman & Bleistein, 1997), and the related concept; item impact. A biased item displays DIF; however that is not sufficient for the item being biased. DIF is a statistical property of an item while item bias is more general and lies in the interpretation (Camilli & Shepard, 1994; Clauser & Mazor, 1998). An item is said to be biased when test takers from one group are less likely to answer an item correctly than test takers of another group due to some characteristic of the item or the test situation that is not relevant to the purpose of the test. If a difference is observed it does not mean that there exists measurement bias since it might be a real difference in ability (Camilli, 2006). For a discussion of item bias see e.g. Penfield & Camilli (2007). Item impact refers to when test takers from different groups have different probabilities of responding correctly to an item due to true differences in ability measured by the item (Dorans & Holland, 1993). Item impact can be measured through the proportion of test takers passing an item regardless of their total score.

One of the major challenges in applying tests is to assure that the tests are fair (see e.g. Camilli, 2006; Gipps, 1994; Shephard, 1982, for a discussion of fairness), in the sense that the most able test takers receive the best test scores. It is not enough to discover DIF items in order to claim unfairness of a test. How fair a test is depends on how the test is used. However, in order to make meaningful comparisons it is required that measurement equivalence holds between different groups of test takers. Since a test is never perfect, a test score can to some degree reflect other variables than those intended to be measured. This is a threat to the validity of inferences drawn from the test scores, which may lead us to consider a test as biased against a group of test takers with certain characteristics. Test bias refers to the systematic difference in total test score against a particular group (Camilli, 2006; Camilli & Shepard, 1994; Wonsuk, 2003). Note, a test free from test bias may contain item bias, if there are about the same amount of items that gives disadvantages to each of two groups they cancel each other out (Hong & Roznowski, 2001). The potential bias in test scores for specific groups such as gender or ethnicity has drawn the attention of both the public and test developers to this matter. There have even been court decisions that have restricted how specific tests for admission decisions should be used since there were evidence that the tests were biased against female and/or minority test takers (Linn & Drasgow, 1987). It is also important to stress that a test needs to be a valid measurement since in some DIF methods the total test score is used as a measure of a test takers ability (Ironson, 1982).

Since DIF analysis was put into light of the measurement industry there has been extensive research and method development for detecting DIF. To examine a measurement quality and its bias one can either use a judgmental and/or a statistical approach, where the latter approach always should be used although the former approach might be handy before testing the items. The test taker group of interest is labeled the *focal group* (usually the minority group) and its performance on an item is compared with the *reference group*. In reality, there may be many pairs of focal and reference groups that DIF is analyzed within (Holland & Wainer, 1993). To detect DIF one can either use an external criterion separated from the test or an internal criterion within the test (Camilli, 2006; Camilli & Shepard, 1994). In this report the focus will be on internal measure criteria within a test. It must be stressed that it is not enough to use a statistical test to detect DIF in an item. We also need to measure the size of DIF.

1.1 Previous DIF studies

Item bias research goes back almost a century starting with Alfred Binet in 1910 who removed some items from a test since it relied to heavy on factors such as scholastic exercise, attention, language (see e.g. Camilli (1993) or Camilli & Shephard, (1994) for details on historically development of DIF research). Also, Angoff (1993) and Cole (1993) described the history and development that has followed the emerging of methods for detecting DIF. More recently, Camilli (2006) have discussed DIF in the context of test fairness.

There have been a number of reviews of methods to detect or measure DIF, see e.g. Berk (1982), Clauser & Mazor, (1998), Camilli & Shephard (1994), Shephard, Camilli & Williams (1985), Holland & Wainer (1993), Millsap & Everson (1993), Camilli (2006) or Penfield & Camilli (2007). There have also been a number of model comparisons for detecting or measuring the size of DIF (see e.g. Hidalgo & López-Pina, 2004; Rogers & Swaminathan, 1993; Wonsuk, 2003; Zumbo, 1999). This report is different from the others mentioned. The focus is on reviewing suitable methods to be use with a criterion-referenced licensing test and comparing them with each other and with respect to some chosen criteria. This is conducted from a theoretical perspective. The main difference when using a criterion-referenced test as compared to a norm-referenced test is the use of a cut-off score, i.e. when a test taker is considered a master or not. Methods which can handle the cutoff score are therefore of special interest here.

1.3. Limitations of the report

DIF methods for use in test-lets, multiple measures, DIF in multiple groups (Kim, Cohen, & Park, 1995), and methods based on expert judgment were not discussed since this was beyond the scope of this report. Further, methods for detecting predictive bias which can be found in educational or personnel selection contexts are not discussed because the test in focus here is not a selection test in the same sense as e.g. a scholastic aptitude test is. The early methods for detecting DIF, which rely on classical test theory (e.g. the delta-plot see Angoff (1993) for a summary) and ANOVA (Rudner, Getson, & Knight, 1980), correlation and reliability estimation methods are also excluded since they confound group difference in test performance on an item with the group difference on average ability. This means that items can be falsely identified as DIF items or items which have DIF are not detected. Camilli & Shephard (1994) and Shephard, Camilli & Williams (1985) emphasize that these methods are inaccurate and should never be used to make judgment about item bias. Further, Dorans & Holland (1993) provide some guidelines to some of these methods. This study only discusses the methods that have been applied to a larger extent.

Specific methods for detecting DIF in polytomously scored items are not discussed although many of the dichotomously scored methods discussed can be extended to use for polytomously scored items. For example; logistic regression (Rogers & Swaminathan, 1993; Wang & Lane, 1996), the SIBTEST, Mantel-Haenszel test and all the IRT methods are methods that can be used or extended for this purpose. Methods such as the logistic discriminant function (Miller & Spray, 1993; Millsap & Everson, 1993; Wang & Lane, 1996) can be used for dichotomously scored items but is particular useful for polytomously scored items are not included here. Further, structural equation modeling which are especially useful for multi-group analysis (Fleishman, Spector, & Altman, 2002) are not included here. For a review of detecting DIF in polytomously scored items see e.g. Penfield & Camilli (2007).

1.2 Aim

The overall aim is to review and compare methods for detecting and measuring DIF in a dichotomously scored criterion-referenced licensing test which is one dimensional, as e.g. the Swedish theory driving-license test. Six criteria are chosen for this purpose and the methods are discussed with respect to them. The different criteria were; parametric or nonparametric, the nature of the matching variable, if they can handle both dichotomously or polytomously scored items, if they can detect and/or measure size of DIF and if they can detect uniform or non-uniform DIF. The methods are also discussed as to whether they can handle the cut-off score specifically. Finally sample size requirements are discussed.

2. Method

Several commonly used DIF methods will be described, compared and categorized in accordance to six criteria. These criteria have been chosen to fit the aim of this study. Some of the criteria have been used in other studies although the last criterion is unique for this study. The first criterion is whether the methods are *parametric or non-parametric*, i.e. whether the model of the item is in focus or the data material is in focus. Within each of these two groups it is also possible to make a distinction between those methods which are contingency tables approaches and those who are not, however this was not used as a criterion. Note that both the non-parametric Mantel-Haenszel procedures and the parametric logistic regression are included in the contingency table approaches. The second criterion is whether the *matching variable* is based on an observed (e.g. total test score) or a latent variable. The first and second criteria are in line with Potenza & Dorans (1995) classification scheme. Although we are primarily interested in methods for *dichotomously* scored items, the third criterion is whether the method can handle or be extended for use with polytomously scored items. This criterion is included for the sake of illustrating flexibility and generalizability of the method. The fourth criterion is between if a DIF methods can measure the effect size of DIF and test DIF, i.e. *measure and/or test DIF*. This criterion is included since the method shows complexity and flexibility if this criterion is fulfilled.

The fifth criterion includes which kind of DIF the methods can handle; i.e. *uniform and/or non-uniform DIF*. An item displays uniform DIF if there is no interaction between ability level and group membership, i.e. the probability of answering an item correctly is greater for one group uniformly over all matched ability levels. An item displays non-uniform DIF if there is an interaction between ability level and group membership. For an item which displays non-uniform DIF the probability of answering an item correctly is not the same over all matched ability levels for the groups (Mellenbergh, 1982). Although Hanson (1998) claims that one also should make a distinction between unidirectional and parallel DIF instead of only reporting uniform DIF they have not been separated in this report because they are only relevant for the item response functions. The sixth and final criterion is whether the method can *handle the cut-off score* in a special way or not. This criterion is included since the focus is on criterion-referenced licensing test and hence we should pay special attention to the cut-off score. Finally, a discussion of required sample sizes was included.

2.1 A criterion-referenced licensing test

An example of a criterion-referenced licensing test is the Swedish theory driving-license test. The test takers have to answer at least 52 (i.e. the cut-off score) items correctly out of the 65 dichotomously scored multiple-choice items in order to pass the test (SRA, 1996). Several different groups of test takers are exposed to the test which makes it especially important that test takers with equal ability are not discriminated against depending on which group they belong to e.g. different ethnic or gender groups. DIF has only been part of one study of the Swedish driving license test (Wiberg, 2006). It is, however, common to study DIF in standardized Swedish tests, see e.g. Stage (1999) or Wester (1997) who examined DIF in the Swedish Scholastic Assessment Test

3. Theoretical review of DIF methods

3.1 Non-parametric methods

Non-parametric methods are not based on a specific statistical model although they may rely on strong assumptions. These methods are particular useful when the sample sizes are small for the groups of interest (Camilli, 2006). In this section contingency table approaches and the SIBTEST will be described. Since Mantel-Haenszel procedures are the most well-known and used procedure it will be treated separately although it belongs to the contingency table approaches.

3.1.1 Mantel-Haenszel procedures

Mantel-Haenszel procedures belong to contingency tables procedures, together with logistic regression, log linear models and some simpler indices. Mantel and Haenszel DIF procedures developed from Mantel &

Haenszel (1959) and was proposed as a method for detecting DIF by Holland & Thayer (1988). Although these methods can easily be extended to polytomously scored items only the dichotomously scored item approach will be described here. The Mantel-Haenszel (MH) test is the most often used method for detecting uniform DIF (Clauser & Mazor, 1998).

The MH test relies on K contingency tables for each item as is seen in Table 1, where each group of test takers are compared regarding their result on the item given their total test scores, i.e. a $2 \times 2 \times K$ table, where $K = 1..., n-1$ is the various total test scores except for 0 and n (Holland & Thayer, 1988).

Table 1. Contingency table for an item for the reference and focal group with total test score k .

	Item score		
	Correct = 1	**Incorrect = 0**	**Total**
Reference group	n_{11k}	n_{12k}	n_{1+k}
Focal group	n_{21k}	n_{22k}	n_{2+k}
Total	n_{+1k}	n_{+2k}	n_{++k}

The MH test statistic tests the null hypothesis of no relation between group membership and test performance on an item after controlling for ability (usually in terms of overall test performance). The test is based on the odds ratio between correct and incorrect responses, between a reference and a focal group when conditioning on total test score. In general, the odds ratio is defined as

$$OR_{RF} = \frac{\pi_R(1-\pi_F)}{\pi_F(1-\pi_R)}$$

where π_F and π_R represent the probability to answer an item correct for the focal and reference group respectively. If the odds ratio is 1 it means that there is no difference between the focal group and the reference group. The resulting test statistic has a chi-squared distribution with one degree of freedom,

$$\chi^2_{MH} = \frac{\left[\left|\sum_{k=1}^{K} n_{11k} - E[n_{11k}] - 1/2\right|\right]^2}{\sum_{k=1}^{K} Var[n_{11k}]},$$

where the expected value is $E[n_{11k}] = (n_{1+k} n_{+1k}) / n_{++k}$, the variance is $Var[n_{11k}] = [n_{1+k} n_{2+k} n_{+1k} n_{+2k}] / n^2_{++k}(n_{++k} - 1)$. ½ is the Yates continuity correction which is used because this is based on a normal approximation of the uniformly most powerful unbiased test and therefore the odds ratio is 1 (Cox, 1988). The odds ratio is required to be uniform and n_{11k} must follows a hypergeometric distribution. A variation of MH test is the Mantel-Haenszel-Cochran test which basically yields the same result but the continuity correction is removed and the variance is slightly changed.

As a measure of the effect size the estimate of the constant odds ratio is used on item level and has range $(0, \infty)$

$$\alpha_{MH} = \frac{\sum_{k=1}^{K} n_{11k} n_{22k} / n_{++k}}{\sum_{k=1}^{K} n_{12k} n_{21k} / n_{++k}},$$

This measure is usually transformed into

$$\beta_{MH} = \ln \hat{\alpha}_{MH}. \qquad (1)$$

A positive value of (1) indicates DIF in favor of the reference group, while a negative value indicates DIF in favor of the focal group. For historical reasons, (1) is transformed into

$$MH\ D - DIF = -2.35 \ln \hat{\alpha} \qquad (2)$$

(Angoff, 1993; Holland & Thayer, 1985). A negative value of (2) indicates that the item is more difficult for the focal group. The Educational Testing Service, ETS, categorizes the degree of DIF in the items as follows (Zieky, 1993).

A_{MH} Negligible DIF, items have $MH\ D - DIF$ not significantly different from zero using χ^2_{MH} or $|MH\ D - DIF| < 1$

B_{MH} Intermediate DIF, items have $MH\ D-DIF$ significantly different from zero and either (a) $|MH\ D-DIF|<1.5$ or (b) $|MH\ D-DIF|$ above 1, but not significantly different from 1.

C_{MH} Large DIF, items have $MH\ D-DIF$ significantly greater than 1.0 and $|MH\ D-DIF|\geq 1.5$.

Longford, Holland & Thayer (1993) comments that if an item is classified as A one can still include the item. If the item is classified as B one should examine if there are other items one can choose to include in the test instead, i.e. an item with a smaller absolute value of $MH\ D-DIF$. Finally, an item classified as C should only be chosen if it meets essential specifications but documentation and corroboration by a reviewer is required. It should also be noted that the number of test takers in the focal group can have a strong influence on the DIF categorization, i.e. more items are classified as category B and C with larger focal and reference group sizes.

3.1.2. Non-parametric contingency table approaches

There are other non-parametric contingency table methods than MH that have been used for measuring the effect size of DIF. One of them is the *proportion difference measure* which is also referred to as *standardization* (Dorans & Kulick, 1986). The idea is to combine the difference in proportion of test takers who answer an item correctly across the focal and reference group given their levels of total test scores. They use a weighted average of the difference in proportions between the two groups that accounts for the number of test takers on each level of total test score. There are two versions; the *unsigned proportion difference* and the *signed proportion difference* indices depending on whether one takes into account the sign of the difference or not. They are also referred to as *standardized p-differences* and *root-mean weighted squared differences* respectively. See e.g. Dorans & Holland (1993) or Camilli & Shephard (1994) for a description. The most commonly used is the *standardized p-difference* which is defined as

$$STD\ P-DIF=\frac{\sum_{k=1}^{K}n_{2+k}\left(\frac{n_{11k}}{n_{1+k}}-\frac{n_{21k}}{n_{2+k}}\right)}{\sum_{k=1}^{K}n_{2+k}}$$

The item is interpreted as a DIF item if the difference is either >0.10 or <-0.10 (Dorans, 1989). This measure is highly correlated with $MH\ D-DIF$ across items (Camilli & Shepard, 1994; Donoghue, Holland, & Thayer, 1993). Also, as with MH it uses observed test score as matching variable. Zieky (1993) denotes that ETS uses both the MH test and the standardized p-difference in their routines since they are easy to work with and give stable results. The latter is especially useful for measuring the size of DIF. Camilli & Shephard (1994) also recommend using the standardized difference index since it is a good description and can be used for explaining the nature of DIF. An advantage with standardization is that it is simple although it lacks a test of significance (Clauser & Mazor, 1998; Millsap & Everson, 1993).

Other non-parametric contingency methods for testing for DIF than MH are closely connected to the indices of measuring DIF. These are later referred to as chi-square methods. For example, the null hypothesis is either that the proportion correct between the reference and the focal group is the same or that their odds ratio is 1. The *summed chi-square* for identifying DIF is described in Camilli & Shephard (1994) but the original test can be traced back to Fisher (1938). The basic idea is to calculate a chi-square statistic on each ability level in a 2×2 table and then combine them into one test statistic for all ability levels. This statistic detects any departure that is large enough but is not commonly used since e.g. the MH is a more powerful nonparametric method (Camilli & Shephard, 1994; Holland & Thayer, 1988). Methods that are no longer in use include a chi-squared method developed by Scheuneman (1979) which was criticized by Baker (1981) for yielding values that were irrelevantly affected by the size of the sample and with no known sampling distribution, meaning not a chi-squared test. Angoff (1993) also describes a full chi-square procedure that was used before the development of MH. The chi-square tests have the advantage of being reliable within usual standards and being homogeneous (Ironson, 1982).

3.1.3 SIBTEST

The Simultaneous Bias Test, SIBTEST, was proposed by Shealy & Stout (1993) and is a modification of the standardization index to detect DIF. In the SIBTEST one tests the null hypothesis

$$H_0 : B(T) = \pi_R(\tau) - \pi_F(\tau) = 0$$

where *B(T)* is the difference in probability (π) of correct response between the reference (R) and the focal group (F) on a specific item when matched on true score τ . The test statistic is

$$\hat{B} = \frac{B_U}{\hat{\sigma}(\hat{B}_U)},$$

where $\hat{B}_U = \sum_k \hat{p}_k (\bar{Y}^*_{R_k} - \bar{Y}^*_{F_k})$, i.e. the average weighted item difficulty difference when one has controlled for the matching variable. $\hat{\sigma}(\hat{B}_U)$ is the standard deviation and the test statistic $\hat{B}$ is normally distributed (Wonsuk, 2003). $\hat{p}_k$ is the proportion of test takers in the focal group who obtained the score $X = k$ on the subtest; $\bar{Y}^*_{R_k}$ and $\bar{Y}^*_{F_k}$ are the adjusted means in the subgroup *k* for the test takers when a regression correction procedure is used. (Shealy & Stout, 1993). The regression correction procedure controls the effects of type 1 error in the valid subtest items (Grierl, Khaliq, & Boughton, 1999; Stout, 2002). It is also possible to measure the amount of DIF in the SIBTEST using the estimate $\hat{\beta}$. Because this estimate is highly correlated with $MH\ D - DIF$ Roussos & Stout (1996) proposed the following guidelines to evaluate the size of DIF, and these have also been used by Zheng, Gierl & Cui (2007)

A_β Negligible DIF, $|\beta| < 0.059$ and H_0 is rejected.

B_β Moderate DIF, $0.059 \le |\beta| \le 0.088$ and H_0 is rejected.

C_β Large DIF, $|\beta| > 0.088$ and H_0 is rejected.

The SIBTEST was initially used to study groups of items simultaneously, although it is also possible to study one item at the time (Camilli, 2006; Narayanan & Swaminathan, 1996; Roussos & Stout, 1996; Shealy & Stout, 1993). The SIBTEST performs well compared with MH when only analyzing one item. The method was developed to model multidimensional data but it is suitable for one-dimensional data as well. The latent ability space is viewed as multidimensional, and includes a one dimensional ability of interest together with nuisance ability. In the beginning all items are used in the matching criterion, but if an item displays DIF it is removed from the matching criterion. The process is repeated until a valid subset of items is identified that does not contain any DIF items. The items can also be divided into two subsets depending on whether they are suspected to have DIF or not. A disadvantage of the SIBTEST is that it is not entirely robust to between-group differences in the unconditional distribution of the ability. This is a problem which the SIBTEST shares with other methods which use the proportion difference, if between group effects are present in the unconditional distribution of the examinees ability (Penfield & Camilli, 2007) A problem with using the SIBTEST is that a special computer program is needed. The SIBTEST have high power in detection of non-uniform DIF, and has high agreement in this sense with logistic regression (Narayanan & Swaminathan, 1996).

3.2 Parametric methods

Parametric methods use a specified model to examine DIF. Methods described here include logistic regression, item response theory models, log linear models and mixed effect models. The likelihood ratio test will be treated separately as it can be applied on several different kinds of models.

3.2.1 Logistic regression

Logistic regression (LR) for detecting DIF was first proposed by Swaminathan & Rogers (1990) but it is a well known statistical procedure. LR relies on the following assumptions. First, the dependent variable must be a discrete random variable. Second, there should be a linear relationship between the continuous variables and the dependent variables logit transformation. Third, a test takers' answer on one item should be independent of the test takers' answer on any other items. Fourth, each independent variable should be measured without an error. Fifth, the errors should be uncorrelated with the independent variable, have a mean of zero and be normally distributed. Sixth, the error variance should be constant across levels of the independent variable (homoscedasticity).

Together with these assumptions there should neither exist multicolinearity in the data material nor any influential outliers (Tabachnick & Fidell, 2001).

LR for detecting DIF is based on modeling the probability of answering an item correctly by group membership and a conditioning variable, usually the observed total test score. The presence of DIF is determined by testing the improvement in model fit when the group membership variable and the interaction between test score and group membership variable is added to the model. According to Camilli & Shephard (1994) LR also belongs to contingency table approaches. Let U = 1 if the test taker has answered the item correctly and 0 otherwise. The LR model for the probability that test takers answer an item correctly can be defined as

$$\pi(U = 1) = \frac{e^z}{1 + e^z} \ ,$$

where $\mathrm{z} = \mathrm{logit}(\pi_{ij}) = \ln\lfloor \pi_{ij} / 1 - \pi_{ij} \rfloor$is given from the logit transformation. The three models of interests are

1. $Z = \beta_0 + \beta_1\theta + \beta_2 G + \beta_3(\theta G)$
2. $Z = \beta_0 + \beta_1\theta + \beta_2 G$
3. $Z = \beta_0 + \beta_1\theta$

where θ is the test takers ability (usually represented by the total test score) and G is the group membership, coded as 1 if the test taker is a member of the focal group and 2 if the test taker is a member of the reference group (Swaminathan & Rogers, 1990). Since the coefficients are estimated using maximum likelihood estimation we can test for DIF using likelihood ratio test statistics. The first model is the augmented model and can be used to test for both uniform and non-uniform DIF simultaneously. The second model allows us to test for uniform DIF. The third (null) model is used when there is no DIF in the item. To test if the item has uniform and/or non-uniform DIF we can compare the fit of the augmented model with the null model. If β_2 is significantly separated from zero, it means that the odds of answering an item is different between the two groups. If β_1 is significantly separated from zero it means that the odds of answering an item correctly increase with increased total test score. If β_3 is significantly separated from zero it means that there is non-uniform DIF (Camilli, 2006; Camilli & Shepard, 1994).

The idea is to choose among these three models the model that fits the data best according to a parsimony principle. In the first step model 1 is tested against model 2 using a likelihood ratio test with one degree of freedom. If there is a significant difference we have non-uniform DIF, if it is not significant we proceed to the next step. In the second step, we test model 2 against model 3. If there is a significant difference we have uniform DIF in the item, but if the difference is not significant we conclude that the item does not display DIF.

There is not just one single method for measuring the size of DIF in LR, instead many different methods have been suggested and the two most commonly used are described here. First, Nagelkerke is defined as

$$R_{Nk}^2 = R_{CS}^2 / R_{\max}^2 \,,$$

where R_{CS}^2 is Cox & Snell's R^2 and $R_{\max}^2 = 1 - [L(0)]^{n/2}$, where L(0) is the likelihood for model 1 (O´Connel, 2006). Zumbo (1999), however, suggested using a weighted least squares R squared to measure the effect size, i.e. to measure the amount of uniform or non-uniform DIF when LR is used

$$\Delta R^2 = R^2(\text{model}\,1) - R^2(\text{model}\,3)\,.$$

There are at least two system of categorization of DIF when LR is used (Hidalgo & López-Pina, 2004). Zumbo & Thomas (1997) proposed the following categories according to Hidalgo & López-Pina (2004)

A_{LRZ} Negligible DIF: $\Delta R^2 < 0.13$

B_{LRZ} Moderate DIF: $0.13 \le \Delta R^2 \le 0.26$

C_{LRZ} Large DIF: $\Delta R^2 > 0.26$

The categories B_{LRZ} and C_{LRZ} also require the statistical tests to flag DIF. More recently, Jodoin & Gierl (2001) have proposed using the following guidelines since they are more sensitive to detect DIF (Hidalgo & López-Pina, 2004).

A_{LR} Negligible DIF: $\Delta R^2 < 0.035$

B_{LR} Moderate DIF: $0.035 \leq \Delta R^2 \leq 0.070$
C_{LR} Large DIF: $\Delta R^2 > 0.070$

Note that categories B_{LR} and C_{LR} also require that the null hypothesis of no DIF is rejected. Which categorization to use is left to the reader to decide. An advantage with LR models is their flexibility to include other variables, and other estimates of ability than total test score. The flexible LR model also allows for conditioning simultaneously on multiple abilities (Clauser & Mazor, 1998; Millsap & Everson, 1993) and can be extended to multiple test taker groups (Agresti, 2002; Miller & Spray, 1993). LR has high power in detection of non-uniform DIF, as mentioned earlier it also displays a high agreement with the SIBTEST in this sense (Narayanan & Swaminathan, 1996).

3.2.2. Likelihood ratio test

The Likelihood ratio test (LRT) is used in connection with several models, e.g. both the LR and item response theory models, and is therefore given a special subsection. LRT is based on the idea that item parameters should be invariant across different subpopulations. An item has DIF if the likelihood is different between a (c)ompact model with few parameters (i.e. the parameters are constrained to be the same) and an (a)ugmented model with all variables of interest (i.e. the parameters are allowed to differ). Regardless of model, anchor items are used to define a common latent metric against which the item with suspected DIF can be examined. The anchor items are assumed to be invariant across groups. The idea is to compare the likelihood of two models and choose the model which has the largest likelihood. The LRT test statistic is defined as

$$G^2 = -2\ln\frac{L(\text{model}\,a)}{L(\text{model}\,c)} = -2\left[\ell(c) - \ell(a)\right] \sim \chi^2_{(m)}$$

where *m* is the difference in number of parameters between the augmented and the compact model. The goal is to test if the additional variables in the augmented model are significantly different from zero. This test statistic is distributed as chi-squared with *m* degrees of freedom under the null hypothesis. (Camilli, 2006; Rao, 1973; Thissen, Steinberg, & Wainer, 1988).

It is usually recommended that the overall goodness-of-fit of the model is tested before proceeding with testing for DIF. However, if the number of items in the test are large and there are many observed zeros in the cross-classifying table (according to the item responses) the general multinomial alternative hypothesis is unreasonable and therefore no satisfactory goodness-of-fit test is available at the moment (Thissen, Steinberg, & Wainer, 1993). A second use of LRT is to test for DIF in a specific item. The idea is to first compute the maximum likelihood (ML) estimates of the parameters of the compact model in the specific item and then the likelihood. Then compute the ML estimates and likelihood for the augmented model which includes parameters representing the difference between the reference and the focal group. In the last step the LRT is used to examine the item for DIF (Thissen et al., 1993).

A large weight of evidence so far supports the use of the LRT over other methods (Millsap & Everson, 1993). It is the best measure of statistical significance but not a good effect size index (Wonsuk, 2003). Therefore, it is recommended that graphs are made to inspect the differences if there are any differences. Note, it might be problematic to use LRT if the sample size is small in any of the focal groups (Camilli, 2006).

3.2.3 Item response theory methods

There are a number of Item Response Theory (IRT) methods for detecting DIF due to the fact that there are a number of IRT models (see e.g. Hambleton & Swaminathan (1985) for an introduction to IRT). All IRT methods are parametric methods since they include modeling the items. The most simple and widely used models are the one-, two- and three-parameter logistic models, i.e. 1PL, 2PL and 3PL. IRT relies on the assumptions that the performance on an item can be explained by or predicted from the test takers latent ability and that this relationship can be modeled as an item characteristic curve, ICC. The 3PL model is defined as

$$\pi_i(u_i = 1) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$

where a_i is the item discrimination, b_i is the item difficulty and c_i is the pseudo-guessing parameter. The 2PL model is obtained by setting the pseudo-guessing parameter equal to 0 and the 1PL model is obtained by also setting the item discrimination equal to 1 (Birnbaum, 1968; Hambleton & Swaminathan, 1985; Lord, 1980). All of the methods for IRT use an estimate of the latent ability as a matching variable. The general idea is to estimate the item parameters separately for the reference and the focal group. If the item does not display DIF the item characteristic curve, ICC, should be identical when placed on the same scale. If an item displays DIF it can be shown in different ways. The item might display DIF in only one of the item parameters (e.g. difficulty) or in all of them (see e.g. Camilli & Shephard, 1994 or Clauser & Mazor, 1988). When examining DIF using IRT it is also common to examine Differential Test Functioning, DTF, at the same time. DTF refers to a difference in the test characteristic curves, obtained by summing up the item response functions for each group. Camilli (2006) points out that large sample sizes are needed when using DIF with IRT. It is especially important that the focal group is not too small in order to obtain stable results.

Angoff (1993) pointed out that the 3PL model more accurately model an item because it allows the item discrimination to vary and includes a pseudo-guessing parameter. The 1PL model can be seriously misleading as indicating DIF although it is really just a difference in item discrimination or in guessing (Camilli & Shepard, 1994). Both uniform and non-uniform DIF can be detected using IRT procedures. When using IRT, an item shows uniform DIF when the ICCs for two groups are different but parallel, while an item displays non-uniform DIF when two groups' ICCs are different but none parallel. The area between the two groups' ICC gives a hint of the degree of DIF in the item (Camilli, 2006; Camilli & Shephard, 1994; Swaminathan & Rogers, 1990). There are both statistical tests and measures of effect size of DIF in IRT and they are performed by comparing the item parameters across groups (Camilli & Shepard, 1994; Lord, 1980). There are four general IRT approaches; *General IRT-LR*, *loglinear IRT-LR*, *limited information IRT-LR* and *IRT-D2*. These approaches all give optimal parameter estimates and statistical tests for DIF. The choice between them depends on the data-analytic context. They will be described in short below.

General IRT-LR uses the Bock-Aitkin (Bock & Aitkin, 1981) marginal maximum likelihood estimation algorithm to estimate the parameters and the LRT to examine the significance of observed differences (Thissen et al., 1993). It is easy to perform by using e.g. the program Multilog (Du Toit, 2003). An advantage is that it can be varied in several ways, and it only relies on the assumption that the population distribution is

Gaussian. A disadvantage is that it is labor-intensive and computationally intensive (Thissen et al., 1993).

Limited-information IRT-LR uses generalized least squares estimation for normal-ogive item response models and LRT:s to examine the significance of observed differences (see e.g. Thissen, Steinberg & Wainer (1993). The limited-information IRT-LR uses information in lower order margins of the response-pattern cross-classification of respondents instead of complete table of response pattern frequencies. It is in general computed using LISPCOMP, a program for structural equation modeling (Thissen et al., 1993). This approach will not be discussed further here.

Log linear IRT-LR uses maximum likelihood estimation for log linear item response models and LRT to examine the significance of observed differences (Kelderman, 1990). The largest limitation with the log linear IRT-LR is that it is only suitable when using the 1PL (Rasch) model (Thissen et al., 1993). This method only emphasizes the difference in item difficulty, and disregards any difference in guessing or in item discrimination as opposed to other IRT methods. Hence, it has more limitation than other IRT methods and will not be discussed further in this paper.

IRT-D2 uses marginal maximum likelihood estimation and ratios of parameter estimates for their standard errors to examine the significance of observed differences (see e.g. Thissen, Steinberg & Wainer (1993). IRT-D2 is built on parameter drift, which is a special case of DIF, i.e. if the item parameters (and therefore the ICC:s) differ across groups during time. This method only emphasizes the difference in item difficulty as opposed to other IRT methods, hence it is a more limited method and will not be discussed further in this paper. The interested reader is referred to e.g. Thissen, Steinberg & Wainer (1993) for more details.

3.2.3.1 Testing DIF with IRT

There are at least five IRT methods for testing the statistical hypothesis of no DIF in an item; *test of b difference*, *item drift method*, *Lord's chi-square*, *empirical sampling distributions for DIF indices* and *measurement of model comparisons*. The simplest way to test for DIF is to test the difference of the item difficulty *b* using an ordinary statistical test for testing the null hypothesis of no difference between the two groups of interest (see Lord, 1980). Another, relatively simple, way is to examine the item drift, i.e. to examine any change in difficulty between two items. The idea is to let the reference group and the focal group take the test on different occasions. Any change in the item difficulty is interpreted as a difference between the two groups. The method has practical advantages since it is easy to apply, however it treats item discrimination as equal across group and this may lead to confounding with group differences (Camilli & Shephard, 1994). Since other IRT methods can handle this, these methods are excluded from further discussion in this paper.

Lord's chi-squared test is an extension of the *test of b difference* which also includes differences in item discrimination. Start by constructing the vector of item parameter differences;

$$V = (\hat{a}_F - \hat{a}_R, \hat{b}_F - \hat{b}_R)\,.$$

The test statistic is defined as

$$Q = VS^{-1}V$$

where S is the variance-covariance matrix of differences between the item parameters. Q follows a chi-square distribution with degrees of freedom equal to number of parameter estimated (Camilli, 2006; Lord, 1980). Refer to Lord (1980) for more computational details. Although the method is sensitive to both uniform and non-uniform DIF it has a large disadvantage. It is possible that the null hypothesis of no DIF is rejected although the ICC:s of two groups are similar, because different combinations of item discrimination and item difficulty may produce similar ICC:s although the item does not display DIF (Camilli & Shephard, 1994). Results from the Lord's chi-squared test correlates fairly well with unsigned area indices (Millsap & Everson, 1993; Shepard, Camilli, & Williams, 1984). A disadvantage is that it sometimes rejects the null hypothesis if the unsigned area between two ICC:s is fairly small throughout the range of ability in which most data appear (Millsap & Everson, 1993).

The *empirical sampling distributions for DIF indices* include a number of more or less sophisticated methods. One of the simpler methods is described by Shephard, Camilli & Williams (1984). The idea is to randomly assign test takers to a reference and a focal group and then examine their responses on an item as compared to the rest of the items. The extreme values from the reference and the focal groups are then used as critical values for DIF. Although this method is tempting it is labor-intensive and is usually restricted to methodology studies (Camilli & Shephard, 1994). This method is therefore excluded from further discussion in this paper since the issue of interest here is to find one or more suitable methods to use in practice with empirical data.

The last category *measurement of model comparisons*, labeled *IRT LRT* in later discussions is one of the most widely used and uses the LRT to test if an item has DIF. Refer to the LRT section for a general description of this method. In IRT the compact model is the IRT model tested and the augmented model is the general multinomial model including all possible parameters that could augment the compact model, given the result that the observed and expected frequencies in each cell are equal (Thissen et al., 1993). When using LRT to test for DIF in IRT models the compact model is defined from the constraint that the item's parameters for the two groups of interest are identical. There is only one restricted model in a test but as many augmented models as there are items. For example if we use the 3PL model the augmented model will contain six variables (two variables for each item parameter since there are two groups) and the restricted model will contain three variables (assuming that the item parameters are alike across groups) (Camilli & Shephard, 1994).

3.2.3.2 Measure size of DIF with IRT

There are at least four different IRT measurements of DIF; *simple area indices*, *probability difference indices*, *b parameter difference* and *ICC method for small samples*. *Simple area indices* are descriptive measurements of the area between the ICC for two groups. A small area indicates small DIF and a large area indicate large DIF. It is a simple way to visualize DIF between two groups by comparing the area between the ICC from the two groups, see e.g. Raju (1988) or (1990). However, a disadvantage with the methods is that they may not take into account the region of the ability continuum with the highest density of test takers and the integrals used to estimate the area do not yield finite values if the *c* parameter, in the 3PL model, is not equal across groups (Camilli & Shephard, 1994). It is also problematic to use with polytomously scored items since one has to compare so many different curves. Further, not all simple area indices provide a standard error, although Raju's (1988) method is an exception (Millsap & Everson, 1993). In order to solve the problem with ability continuum in the simple area indices *probability difference indices* have been developed by Linn & Harnisch (1981) and Shephard, Camilli & Williams (1984). The idea is to weight the areas in order to reflect the reliability of the difference between the two ICC:s. The third IRT measurement of DIF; *b parameter indices* is simply the difference in the item difficulty parameter between the two groups (Camilli & Shephard, 1994). The last method of measurement was proposed by Linn & Harnisch (1981) and the basic idea is to compare the item ability parameter estimates for the whole group with the estimates in the focal group. The method has mainly been used in the past when computer power was low and it was difficult to estimate the item parameters. Camilli & Shephard (1994) suggest that *probability differences indices* or *b parameter differences* should be used instead of this last method since they are more reliable. Since the other two methods are more limiting and less reliable they are excluded from further discussion in this study.

The IRT methods are not equally sensitive for DIF. Strong IRT models (e.g. the 1PL model) have the most sensitive tests of DIF when these models are accurate. These models can also compensate to some extent for incomplete data. The signed and unsigned probability difference statistics are easy to calculate and are recommended since they are stable, and can detect DIF in regions where the data occurs in the ICC graph. The model comparison approach is recommended by Camilli & Shephard, 1994). IRT DIF indices based on joint maximum likelihood item statistics should be avoided because the estimation might be poor. It is also possible to test the improvement of the fit of a model by comparing fit between parameter estimates of the whole group compared with parameter estimates of the fit when a specific group has been excluded (Thissen et al., 1993). The IRT LRT tests have been shown to be stable and are easy to extend to permit simultaneous test of bias for multiple items. The only disadvantage is that unbiased anchor items are needed (Millsap & Everson, 1993).

3.2.4 Log linear models

Log linear models (LLM) for detecting DIF were suggested by Mellenbergh (1982) and have been used in the past (see e.g. Kelderman & Macready, 1990)). However, the method is quite powerful and should not be disregarded. It is a parametric method and also a contingency table method which uses the observed test score as matching variable. The general idea is to model the items on accordance with a LLM which accounts for the total test score divided into intervals, a group term and the item difficulty

$$\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(j)} + \mu_{13(k)} + \mu_{23(k)} + \mu_{123(ijk)}$$

where F_{ijk} is the expected frequency, and $\mu_{1(i)}$ is the item score effect, $\mu_{2(j)}$ is the *j*:th group effect and $\mu_{3(k)}$ is the *k*:th score level. This model is usually reformulated as

$$\ln(F_{1jk} / F_{0jk}) = \alpha + \beta_k + \gamma_j + (\beta\gamma)_{jk}$$

where F_{1jk} and F_{0jk} is the expected frequency of correct and incorrect response on item *j* by group *j* on test score level *k* respectively. α is the overall item difficulty, β is the test score effect and γ is the group effect. If this model holds the item displays non-uniform DIF and one can eliminate only the interaction term the item displays uniform DIF and if one can eliminate both group terms the item does not display DIF. To test the model for DIF the LRT, as described previously, is usually used although Pearson's chi-squared test may also be used (Kelderman & Macready, 1990; Millsap & Everson, 1993). Advantages of using LLM include their flexibility and that they can easily be extended to polytomously scored items, multiple test takers or simultaneous DIF detection in several items (Agresti, 2002). However, it has been argued that LLM do not represent data adequately except when the items can be modeled with a 1PL model, because in that model the total score is a sufficient statistic. Millsap & Everson (1993) claim that LLM are less suitable for more complex models such as the 2PL model

3.2.5. *Mixed Effect Models*

A new parametric approach of modeling items has been suggested during the last couple of years, although the statistical technique has existed longer, se e.g. Pinheiro & Bates (2001). The main idea is to view not all factors as fixed but instead view one or more as random. It is most common to view the item parameters as fixed and considered the test takers' parameters as random effects. Mixed-effect models are suitable both for linear and non-linear modeling. Using this approach there are two possibilities for examining DIF; either examine a random item effect or a random group effect (De Boeck & Wilson, 2004). One can study both uniform and non-uniform DIF with this approach for both dichotomously and polytomously scored items using different variations of the mixed effect models. Although the mixed effect modeling approach is flexible a disadvantage is that it is a new method that requires more research. Software development has been made but it has not been used extensively on large data sets and it is computer intensive if the model is complicated. For more information of how to examine DIF using these models see e.g. Meulders & Xie (2004).

4. Comparisons of methods

In order to choose which DIF methods to use the criteria given in the method section will be discussed together with a note on required sample sizes. The comparison focused on criterion-referenced licensing tests with dichotomously scored items that measure one dimensional ability. Only the general methods are discussed and not the more specialized mentioned in the previous section. No matter which method is chosen it is desirable that the method has high statistical power to detect DIF, i.e. high probability of identifying DIF in an item, while controlling for type 1 error, which is. the probability of identifying an item as DIF when the item has no DIF. A summary of results of the comparison with respect to the previously described criteria is given in Table 2. Note, whether the method can take special care of the cut-off score is not included in the table but will nevertheless be discussed, as will sample size requirements.

Table 2. DIF methods categorized depending on their nature. 1. (Par)ametric or (non-p)arametric. 2. Matching variable; (Obs)erved or (Lat)ent 3. D)ichotomusly or (P)olytomously scored items. 4. Whether one can (T)est and/or (M)easure DIF. 5. Able to handle (U)niform and (N)onuniform DIF.

Method	**Par/** **Non-p**	**Obs/** **Latent**	**Item** **scores**	**T/M**	**U/N**
Mantel-Haenszel	Non-p	Obs	D/P	T/M	U
Standardization	Non-p	Obs	D	M	U
Chi-square methods	Non-p	Obs	D	T	U
SIBTEST	Non-p	Lat	D/P	T/M	U/N
Logistic Regression	Par	Obs	D/P	T/M	U/N
Likelihood ratio test	Par	Obs/Lat	D/P	T/M	U/N
Prob. diff. indices	Par	Lat	D	M	U/N
b parameter indices	Par	Lat	D	M	U/N
General IRT-LR	Par	Lat	D/P	T/M	U/N
IRT LRT	Par	Lat	D/P	T	U/N
IRT methods	Par	Lat	D/P	T/M	U/N
Lord's chi-squared test	Par	Lat	D	T	U/N
Log linear models	Par	Obs	D/P	T	U/N
Mixed effect models	Par	Lat	D/P	T	U/N

4.1 Parametric vs. non-parametric

First, if we use a parametric method it is very important that the model assumptions are fulfilled. If the chosen model's assumptions are violated either another model should be chosen or a non-parametric method should be used instead. It is of course also possible to use a non-parametric approach to start with but then it is usually more difficult to control for covariates. An advantage with non-parametric tests, as e.g. the chi-square tests, is the lack of assumption about the distribution of ability in the population of interest (Ironson, 1982). The SIBTEST only assumes monotonicity for example.

All parametric methods have more or less strict assumptions which have to be fulfilled otherwise they should not be used. LR, e.g. rely on the strong assumption that the relationship between the probability of answer an item correct and the observed test score is linear. Both Embretson & Reise (2000) and Lord (1980) have shown that the observed test score is nonlinearly related to the examinees latent ability. MH (Camilli, 2006), Also, using either of the IRT methods means to rely on the strong assumption of one dimension in the test. In conclusion, if the model fits the data and the assumptions are fulfilled a parametric DIF method can be used. If one cannot find a model with model assumptions fulfilled or the model does not fit the data a non-parametric method should be chosen instead. The contingency table approaches (MH, LR, LLM, chi-square tests), i.e. including both non-parametric and parametric approaches, have the disadvantage that no parameter is usually available for guessing and discrimination in these as compared with IRT methods. Instead discrimination among items is usually assumed to be equal across items for the focal and reference groups (Camilli, 2006). MH, LR and LLM work best when the data can be modeled with a 1PL model but is more problematic if the data is modeled with a 2PL or a 3PL model (Millsap & Everson, 1993). Note, under some assumptions LLM can yield the same result as LR.

4.2 Matching variable

The second criterion concerned the nature of the matching variable; observed or latent. To choose between these two usually requires an idea of whether to use classic test theory (observed score) or modern test theory (latent score). This may influence how the whole test is examined and/or how the result is reported. Methods that use the observed score as matching variable include MH, standardization, chi-square methods, LR and LLM. All these methods yield poor results if the total score is a poor proxy of the latent ability, e.g. it includes other abilities too. A general disadvantage with all the methods that use the observed test scores is that it makes the strong assumption that ability is adequately represented by the total test score. The total test score is not a perfect measure of a test takers' ability and it requires that the test is valid (Ironson, 1982; Millsap & Everson, 1993). Note, although both LR and MH uses the observed score as matching variable, in LR the continuous test score variable does not need to be classified, hence there might be less errors. The binary LR model can also be generalized to use with ordinal scores (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Moreover, LR is more efficient than to use multiple matching as required by the MH analysis (Mazor, Kanjee, & Clauser, 1995). MH has been noted to both over- and under-estimate DIF according to several factors such as e.g. matching variable (Holland & Thayer, 1988), guessing (Camilli & Penfield, 1997), and when there is a lack of sufficient statistic for matching (Zwick, 1990).

IRT methods, the non-parametric SIBTEST, and mixed effect models do not use the observed score as matching variable, instead they use a latent variable. It has been argued that the latent score is a more precise measure of the ability of the test takers. The problems of using the observed score as matching variable can be solved by removing biased items iteratively and redo the analysis (van der Flier, Mellenbergh, Ader, & Wijn, 1984). Therefore, no matter which method(s) are chosen it is recommended that all methods should be used iteratively (Camilli & Shephard, 1994).

4.3 Dichotomously vs. polytomously

The third criterion, whether the methods can handle dichotomously and polytomously scored items, less important since the main concern has been dichotomously scored items. This criterion is merely used to show upon if the method is flexible if some items in a test are given a different item format. Note that MH, the SIBTEST, LR, LRT, general IRT-LR, LLM and mixed effect models are all flexible in this sense. Originally, the MH was not design to use with polytomously scored items but has been extended to fulfill this purpose (Zwick, Donoghue, & Grimo, 1993). The LR has also been adapted to polytomously scored items (Camilli &

Congdon, 1999) although there have been reports of difficultness to apply LR to polytomously DIF (Wonsuk, 2003).

4.4 Measure and/or test DIF

The fourth criterion, whether the method can both detect and measure DIF, has to be considered. It is of course possible to choose a method which only handles either of these parts if the chosen method is used as a complement to another method which handles the other part. It also depends on the purpose of performing a DIF study. Is the primary aim to identify problematic items or do we want to measure the size of DIF in order to choose which items to disregard? MH, the SIBTEST, LR, LRT, General IRT-LR and IRT methods are all methods which can both test and measure the size of DIF. The other methods have the disadvantage of either only measuring size of DIF (probability difference indices and b parameter indices) or only providing significant test of DIF (chi-square methods, Lord's chi-square test, LLM and mixed effect models).

4.5 Uniform vs. non-uniform DIF

The fifth criterion, whether the methods can handle uniform DIF and non-uniform DIF is quite important, since we cannot assume that the behavior of DIF is linear. Most non-parametric methods can only handle uniform DIF satisfactorily. The MH, e.g. has the disadvantage that it is designed to measure uniform DIF, which means that it is not that sensitive to non-uniform DIF. It has been suggested that MH can be modified in order to use it for detecting non-uniform DIF (Mazor, Clauser, & Hambleton, 1994), but it is still mostly used for detecting uniform DIF. In Table 2 it is classified as uniform DIF since the detection of non-uniform DIF is not yet widely accepted. Hidalgo & Lopéz (2004) states that MH is somewhat more powerful than LR in detecting uniform DIF. However, the LR has a higher statistical power level than MH approaches in detecting non-uniform DIF (Hidalgo & López-Pina, 2004; Miller & Spray, 1993; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Camilli & Shephard (1994) recommend using LR to for detecting DIF. The conclusion is therefore to use LR instead of MH if LR model assumptions are fulfilled, otherwise use MH. If only uniform DIF is of interest MH can of course be used instead of LR. The standardization method and MH have been shown to give similar results (Wonsuk, 2003), which can lead us to conclude that they can be use interchangeably. Standardization (standardized p-difference) also has problems detecting non-uniform DIF (Penfield & Camilli, 2007). The SIBTEST can, however, detect non-uniform DIF satisfactorily even though it is a non-parametric test. All parametric methods in this study can handle both kinds of DIF.

4.6 Handle the cut-off score

The sixth criterion was especially chosen because we have a criterion-referenced licensing test and it concerns whether the methods allow for special consideration of the cut-off score. One possibility is to divide the test takers' score in any of the contingency table approaches according to whether the test takers' score are below or above the cut-off score. In both MH and LR e.g. it is possible to divide the matching variable and control the ability among e.g. test takers which failed the test or only among test takers who passed the test. Another possibility is to use methods where the observed score are always categorized and included as such in the model as e.g. in LLM, and maybe mixed effect models. Note that it is possible to use LRT with several of these methods. To divide the total score into intervals is problematic if a latent variable is used as matching variable, although it might be possible to model the test takers with respect to whether they are above or below the cut-off score in e.g. mixed models. Another possibility is to model each interval of test takers separately using e.g. IRT models, and then compare them. These ideas have not been tried before and more research is needed.

4.7 Sample size

The contingency table approaches (including MH, LR and LLM) have the advantage that they only require small sample sizes, especially in comparison with IRT (Ironson, 1982; Penfield & Camilli, 2007). In particular, when the sample size of the focal group is small it is problematic to use the IRT methods (Camilli, 2006). Camilli & Shephard (1994) recommend using weak (complex) IRT models (e.g. the 3PL model) for research when there is a large enough sample and using the more inexpensive contingency tables approaches in applications. These IRT methods are also computational intensive (Clauser & Mazor, 1998; Millsap & Everson, 1993).

Swaminathan & Gifford (1983) stated 1000 test takers where needed when 20 items were used. Hulin, Lissak, and Drasgow (1982) observed a trade-off between test-length and sample size; when using either at least 1000 test takers and 60 items or 2000 test takers and 30 items it is possible to make accurate parameter estimates using the 3PL model. Note that the accuracy of the DIF analysis is highly dependent on the validity of the chosen IRT model (Camilli & Shephard, 1994). The SIBTEST works well with a fairly small sample, i.e. from 250 test takers although the test should preferably have 20 items or more (Millsap & Everson, 1993). The (un)signed area indices demand large sample sizes but are not known to give stable results and they do not take into respect the examinees distribution across ability; this means that these indices can exaggerate the amount of DIF in a population.

5. Strategies for performing a DIF analysis

5.1. Selection of methods

The test of interest is a criterion-referenced licensing test with dichotomously scored items which limits the range of possible methods. Camilli & Shephard (1994) repeatedly emphasized the importance of using several methods for testing and measuring DIF and therefore more than one method will be recommended. Keeping the chosen criteria and the above recommendation in mind four methods are recommended to examine empirically for detecting and measuring DIF in a dichotomously scored criterion-referenced licensing tests. A summary of the selected methods is given in Table 3.

Table 3. Selected DIF methods categorized with respect to the given criteria. 1. (Par)ametric or (non-p)arametric. 2. Nature of the matching variable; (Obs)erved or (Lat)ent 3. D)ichotomusly or (P)olytomously scored items. 4. Whether one can (T)est and/or (M)easure DIF. 5. Handle (U)niform and (N)onuniform DIF.

Method	**Par/ Non-p**	**Obs/ Latent**	**Item scores**	**T/M**	**U/N**
LR (Logistic Regression)	Par	Obs	D/P	T/M	U/N
IRT methods	Par	Lat	D/P	T/M	U/N
LLM (Log linear models)	Par	Obs	D/P	T	U/N
MH (Mantel-Haenszel)	Non-p	Obs	D/P	T/M	U

First, LR was chosen because it is a flexible method that can detect both uniform and non-uniform DIF. LR usually uses the observed score as matching variable. Therefore, an IRT procedure was as well chosen because they are also known to detect both non-uniform and uniform DIF but rely on a latent score as matching variable. Both these methods can be extended to polytomously scored items and can not only detect but also measure DIF. However, both of these methods rely on strong model assumptions which have to be examined carefully. For the moment it is not known if these assumptions will be met in the test we have in mind. Therefore, two contingency table approaches; LLM and MH were also chosen. LLM also rely on strong assumptions but they are different from the other chosen methods. Furthermore, LLM can handle both dichotomously and polytomously scored items, as well as uniform and non-uniform DIF. A disadvantage is that there is no known scale for interpreting the size of DIF, however, with more research such a scale might appear in the future. Finally, the MH was chosen because it is a non-parametric method which does not rely on strong model assumptions. Further, it can handle both dichomotously and polytomously scored items, it can both measure and test DIF. It cannot, however, satisfactorily detect non-uniform DIF. In all these four methods it is possible to either model the cut-off score or divide the test takers score in different ways to ensure that the cut-off score is taken into special consideration. All these four methods can be used as complements to each other as long as the model assumptions are fulfilled.

When a method has been chosen one needs to choose a sample to perform the DIF analysis. When performing a DIF study the size of the sample depends not only on the method chosen but also on when the DIF analysis is done. If the DIF analysis is made in the test assembly, at least 100 test takers should be included in the smaller group and a total of at least 500 test takers should be included. If the DIF analysis is performed before reporting the total score but after the regular test, at least 200 test takers should be included in the smaller group and 600 test takers in the total group. If examining DIF after reporting total test score at least 500 test takers in the smaller group should be used (Zieky, 1993). Clauser & Mazor (1998) comment that the larger the sample size the more accurate is the tests, especially when using IRT methods.

5.2 DIF analyses strategies

The overall aim was to find suitable methods for detecting and measuring DIF in a dichotomously scored criterion-referenced licensing test as e.g. the Swedish theory driving-license test. Different methods for detecting DIF have been discussed and compared. The methods were classified as being non-parametric or parametric methods, as to whether their matching variable is observed or latent score, whether the items are dichotomously or polytomously scored, and whether they test and/or measure DIF. Further, it was examined whether they can handle both uniform and non-uniform DIF satisfactory and whether they can treat (or model) the groups above or below the cut-off score especially or not. Finally, a short discussion of required sample sizes was added.

The comparison of methods did not single out one method that can be recommended; instead the suggestion was to examine at least four methods; MH, LR, LLM, and an IRT method. To decide between these methods an empirical study is needed, however that is beyond the scope of this study. Instead this section will focus on the discussion of the meaning of DIF and how to proceed if a test contains DIF items. Note, that to ensure that we have a valid test it is not enough to just examine DIF in a test. A whole validation process is needed, see e.g. Haladyna (2006) for suggestions on how to proceed. It is also important to keep in mind(2006) ideas of validation, i.e. in which context the test is used in order for it to be valid. Here a discussion on how to perform a DIF analysis will be in focus.

First of all, it has to be stated that an item should always be examined before it is put into a test so that it is not offensive or demeaning for any member of a group, see Berk (1982) for a summary of these procedures for six test publishers. An updated version is given by Ramsey (1993) who labeled it sensitivity analysis. This analysis is performed in order to a) balance the test b) not foster stereotypes c) not include gender-based or ethnocentric assumptions d) avoid the test being offensive to any test taker e) not contain controversial material which is not demanded by the subject f) avoid elitism or ethnocentrism.

After the sensitivity analysis has been performed one can pretest the item in a smaller group so that all items can be examined for DIF (Longford, Holland & Thayer, 1993). Burton & Burton (1993) noted that screening pretested item for DIF reduced the amount of DIF items in a regular test. The DIF screening did not have a substantial effect on item difficulty, item discrimination or average test scores of the focal group as compared with the reference group. It is also important to emphasize that easy items are more likely to be flagged as displaying DIF than more difficult items (Linn, 1993). Therefore, it is wise to compare the DIF statistics obtained with other item statistics since they are related to overall item difficulty and item discrimination (Burton & Burton, 1993; Linn, 1993). The item discrimination was defined in this case as the biserial correlation between the item and the total test score and the difference in difficulty between the focal and the reference group (Linn, 1993). Item difficulty was defined as percentage correct on the item. In the Burton & Burton (1993) study the MH D-DIF was used as DIF measure.

Note that DIF studies can never be performed routinely without reflection, they have to be followed up with an examination of why a particular item displays DIF. A statistical inference test which gives significant result of DIF does not imply practical significance and needs to be complemented with practical measures of size. If an item in a test displays DIF, one should try to find the source of the DIF, because it is not necessarily a bad item. An item might display DIF if it has a different item format than the rest of the items in the test (Longford et al., 1993). Another possibility is that the item measures an ability different from the one measured in the test or reflect that two groups have learned something with different pedagogical methods, hence making an item easier for one of the groups (Camilli, 2006). If it really is an item that favors one group, conditional on the ability, there are some strategies that one can apply. The most common ones are a) rewrite the item b) remove the item c) control for the underlying differences using an IRT model for scoring respondents. If however the item is kept in the test the test constructor should have a reason for that decision.

If an item displays DIF there also has to be a judgment whether the item is unfair to any of the groups. How fair the item is depends on the purpose of the test. It is also possible that the difference in total test scores reflects a genuine difference in ability between the groups, hence no DIF exists. DIF statistics can only answer trivial questions such as: “Do the items measure the same across different groups?” However, they do not address questions such as the (un)intended consequences of the test or if the test is fair (Camilli & Shepard, 1994; Zieky, 1993). It is therefore important that one decides how large a DIF is reasonable in an item before it is removed. There also has to be guidelines about when an item should be sent for review. Should it be sent only if it favors the reference group (i.e. the majority) or also if it favors the focal group? What action should be done if the item displays DIF? Should it be removed or rewritten? Usually, this depends on the measurement decision or the seriousness of the measurement errors (Zumbo, 1999).

In practice it is rare to remove items which display DIF in a given test, unless it is discovered before the test is administrated. However, examination of DIF can help test developers to construct more fair tests (Penfield & Camilli, 2007). A word of warning, just because a test has no items that display DIF does not mean that the test is fair. As DIF analysis rely on an internal criterion, e.g. using the total test score as the ability, DIF studies cannot detect constant bias. If all items in a test displays DIF we will not be able to detect this because the observed score will in general be underestimated and we are using the estimated ability of the test takers as a control. Another problem when running a DIF analysis is that it might be problematic to define the grouping variable if we chose e.g. educational background (Penfield & Camilli, 2007).

One way of improving the analysis is to use the DIF methods iteratively, i.e. to remove the biased items and re-examine DIF to check for potentially biased items (Camilli & Shephard, 1994). Note, however, that the potential DIF item should always be included in the matching variable test score otherwise the DIF analysis can yield strange results and e.g. the MH procedure does not work correctly if the DIF item is removed (Dorans & Holland, 1993; Holland & Thayer, 1988; Lewis, 1993; Zwick, 1990). Also, it might be impossible to remove all DIF, because the focal and reference group do not have the same life experience.

5.3 Further research

This has been a theoretical review of possible DIF methods to be used with a dichotomously scored criterion-referenced licensing test, as e.g. the Swedish theory driving-license test. The next step is to examine and compare the suggested methods using empirical data to see if they yield similar results and if the assumptions they rely on hold. If an item displays DIF it is also important to examine the item carefully in order to try to explain why the item displays DIF.

6. References

Ackerman, T. A. (1992). A didactive explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Dublin: Educational Research Center.

Baker, F. B. (1981). A criticism of Scheuneman's item bias technique. *Journal of Educational Measurement, 18*, 59-62.

Berk, R. A. (1982). *Handbook of methods for detecting test bias*. London: The John Hopkins Press.

Birnbaum, A. (1968). In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika, 46*, 443-449.

Burton, E., & Burton, N. W. (1993). The effect of item screening on test scores and test characteristics. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 321-335). Hillsdale, NJ: Lawrence Erlbaum Associates.

Camilli, G. (1993). The case against DIF techniques based on internal criteria: Do item bias procedures obscure test fairness? . In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: theory and practice* (pp. 397-417). Hillsdale, NJ: Lawrence Erlbaum.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., Vol. 4, pp. 221-256). Westport: American Council on Education & Praeger Publishers.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics, 24*(4), 323-341.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage publications.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement, 17*(1), 31-44.

Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 25-29). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cox, D. R. (1988). *Analysis of binary data* (2nd ed.). London: Nethuen.

De Boeck, P., & Wilson, M. E. (2004). *Explanatory item response models: a generalized linear and nonlinear approach*. New York: Springer-Verlag.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2*, 217-233.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenzel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dorans, N. J., & Kulick, E. M. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance in the scholastic aptitude test. *Journal of Educational Measurement, 23*, 355-368.

Du Toit, M. (2003). IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact: Scientific software international.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum Associates inc. publishers.

Fischer, R. A. (1938). *Statistical methods for research workers* (4th ed.). London: Oliver & Boyd.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journal of Gerontology: social sciences, 57B*, 275-284.

Gipps, C. (1994). *Beyond testing: towards a theory of educational assessment*. London: Routledge Falmer.

Grierl, M., Khaliq, S. H., & Boughton, K. (1999). *Gender differential item functioning in mathematics and science: prevalence and policy implications.* Paper presented at the symposium entitled Improving large-scale assessment in education at the annual meeting of the Canadian society for the Study of Education.

Haladyna, T. M. (2006). Roles and importance of validity studies in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 739-755). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 23*(3), 244-253.

Hidalgo, M. H., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: a comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*, 903-915.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (No. 80-43). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale NJ: Erlbaum.

Holland, P. W., & Wainer, H. E. (1993). *Differential item functioning*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Hong, S., & Roznowski, M. (2001). An investigation of the influence of internal test bias on regression slope. *Applied Measurement in Education, 14*, 351-368.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*(3), 249-260.

Ironson, G. H. (1982). Use of chi-squared and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias*. London: The John Hopkins Press.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type 1 error and power rates using an effect size measure with logistic regression procedures for DIF detections. *Applied Measurement in Education, 14*, 329-349.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport: American Council on Education & Praeger Publishers.

Kelderman, H. (1990). Item bias detecting using loglinear IRT. *Psychometrika, 54*, 681-697.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27*, 307-327.

Kim, S.-E., Cohen, A.-S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*(3), 261-276.

Lewis, C. (1993). A note on the value of including the studied item in the test score when analyzing test items for DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 321-335). Hillsdale, NJ: Lawrence Erlbaum Associates.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Linn, R. L., & Drasgow, F. (1987). Implications of the golden rule settlement for test construction. *Educational Measurement, 6*, 13-17.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership of achievement test items. *Journal of Educational Measurement, 18*, 109-118.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171-196). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54*, 284-291.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*, 131-144.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-108.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: American Council on Education & Macmillan.

Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: a* *generalized linear and nonlinear approach* (pp. 213-240). New York: Springer.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*(2), 107-122.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334.

Narayanan, & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274.

O´Connel, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks: SAGE.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics Psychometrics* (Vol. 26, pp. 125-167). Amsterdam: Elsevier.

Pinheiro, J., & Bates, D. M. (2001). *Mixed-effects models in S and S-plus*. New York: Springer-Verlag.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: a framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

Raju, N. S. (1988). The area between two item characteristics curves. *Psychometrika, 54*, 495-502.

Raju, N. S. (1990). Determing the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207.

Ramsey, P. A. (1993). Sensitivity review the ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367-388). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rao, C. R. (1973). *Linear statistical inference and its applications*. New York: Wiley.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(105-116).

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type 1 error performance. *Journal of Educational Measurement, 33*, 215-230.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item techniques. *Journal of Educational Statistics, 213-233.*

Scheuneman, J. D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*, 143-152.

Scheuneman, J. D., & Bleistein, C. A. (1997). Item bias. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: an international handbook* (2nd ed., pp. 742-748). New York: Elsevier.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detect test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Shepard, L. A., Camilli, G., & Williams, A. F. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*, 77-105.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artefacts in item bias research. *Journal of Educational Statistics, 9*, 93-128.

Shephard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias*. London: The John Hopkins Press.

SRA. (1996). *VVFS 1996:168 Vägverkets författningssamling. vägverkets föreskrifter om kursplaner, behörighet B. [Regulations concerning driving license class B]*. Borlänge: Swedish Road Administration.

Stage, C. (1999). *Predicting gender differences in word items. A comparison of item response theory and classical test theory* (EM No. 34): Department of Educational Measurement, Umeå University, Sweden.

Stout, W. (2002). Psychometrics: from practice to theory and back. *Psychometrika, 67*(4), 485-518.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 9-30). New York: Academic Press.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston: Allyn and Bacon.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

van der Flier, H., Mellenbergh, G. J., Ader, H. J., & Wijn, M. (1984). An iterative item bias detecting method. *Journal of Educational Measurement, 21*, 131-145.

Wang, N., & Lane, S. (1996). Detection of gender-related differential item functioning in a mathematics performance assessment. *Applied Measurement in Education, 9*(2), 175-199.

Wester, A. (1997). *Differential item functioning (DIF) in relation to item content* (EM No. 27): Department of Educational Measurement, Umeå University, Sweden.

Wiberg, M. (2006). Gender differences in the Swedish driving-license test. *Journal of Safety Research, 37*, 285-291.

Wonsuk, K. (2003). *Development of a differential item functioning (DIF) procedure using the hierarchical generalized linear model: a comparison study with logistic regression procedure.* Pennsylvania State University.

Zheng, Y., Gierl, M. J., & Cui, Y. (2007). Using real data to compare DIF detection and effect size measures among mantel-Haenszel, SIBTEST, and logistic regression procedures *Paper presented at NCME 2007*. Chicago.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): logistic regression modelling as a unitary framework for binary and likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defence.

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF* (Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science). Prince George, Canada, University of British Columbia.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*(185-197).

Zwick, R., Donoghue, J. R., & Grimo, A. (1993). Assessing differential item functioning in performance tasks. *Journal of Educational Measurement, 30*, 233-251.

EDUCATIONAL MEASUREMENT

Reports already published in the series

EM No 1. SELECTION TO HIGHER EDUCATION IN SWEDEN. Ingemar Wedman

EM No 2. PREDICTION OF ACADEMIC SUCCESS IN A PERSPECTIVE OF CRITERION-RELATED AND CONSTRUCT VALIDITY. Widar Henriksson, Ingemar Wedman

EM No 3. ITEM BIAS WITH RESPECT TO GENDER INTERPRETED IN THE LIGHT OF PROBLEM-SOLVING STRATEGIES. Anita Wester

EM No 4. AVERAGE SCHOOL MARKS AND RESULTS ON THE SWESAT. Christina Stage

EM No 5. THE PROBLEM OF REPEATED TEST TAKING AND THE SweSAT. Widar Henriksson

EM No 6. COACHING FOR COMPLEX ITEM FORMATS IN THE SweSAT. Widar Henriksson

EM No 7. GENDER DIFFERENCES ON THE SweSAT. A Review of Studies since 1975. Christina Stage

EM No 8. EFFECTS OF REPEATED TEST TAKING ON THE SWEDISH SCHOLASTIC APTITUDE TEST (SweSAT). Widar Henriksson, Ingemar Wedman

1994

EM No 9. NOTES FROM THE FIRST INTERNATIONAL SweSAT CONFERENCE. May 23 - 25, 1993. Ingemar Wedman, Christina Stage

EM No 10. NOTES FROM THE SECOND INTERNATIONAL SweSAT CONFERENCE. New Orleans, April 2, 1994. Widar Henriksson, Sten Henrysson, Christina Stage, Ingemar Wedman and Anita Wester

EM No 11. USE OF ASSESSMENT OUTCOMES IN SELECTING CANDIDATES FOR SECONDARY AND TERTIARY EDUCATION: A COMPARISON. Christina Stage

EM No 12. GENDER DIFFERENCES IN TESTING. DIF analyses using the Mantel-Haenszel technique on three subtests in the Swedish SAT. Anita Wester

1995

EM No 13. REPEATED TEST TAKING AND THE SweSAT. Widar Henriksson

EM No 14. AMBITIONS AND ATTITUDES TOWARD STUDIES AND STUDY RESULTS. Interviews with students of the Business Administration study program in Umeå, Sweden. Anita Wester

EM No 15. EXPERIENCES WITH THE SWEDISH SCHOLASTIC APTITUDE TEST. Christina Stage

EM No 16. NOTES FROM THE THIRD INTERNATIONAL SweSAT CONFERENCE. Umeå, May 27-30, 1995. Christina Stage, Widar Henriksson

EM No 17. THE COMPLEXITY OF DATA SUFFICIENCY ITEMS. Widar Henriksson

EM No 18. STUDY SUCCESS IN HIGHER EDUCATION. A comparison of students admitted on the basis of GPA and SweSAT-scores with and without credits for work experience. Widar Henriksson, Simon Wolming

1996

EM No 19. AN ATTEMPT TO FIT IRT MODELS TO THE DS SUBTEST IN THE SweSAT. Christina Stage

EM No 20. NOTES FROM THE FOURTH INTERNATIONAL SweSAT CONFERENCE. New York, April 7, 1996. Christina Stage

1997

EM No 21. THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SWESAT. A study of the DTM subtest. Christina Stage

EM No 22. ITEM FORMAT AND GENDER DIFFERENCES IN MATHEMATICS AND SCIENCE. A study on item format and gender differences in performance based on TIMSS´data. Anita Wester, Widar Henriksson

EM No 23. DO MALES AND FEMALES WITH IDENTICAL TEST SCORES SOLVE TEST ITEMS IN THE SAME WAY? Christina Stage

EM No 24. THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SweSAT. A Study of the ERC Subtest. Christina Stage

EM No 25. THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SweSAT. A Study of the READ Subtest. Christina Stage

EM No 26. THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SweSAT. A Study of the WORD Subtest. Christina Stage

EM No 27. DIFFERENTIAL ITEM FUNCTIONING (DIF) IN RELATION TO ITEM CONTENT. A study of three subtests in the SweSAT with focus on gender. Anita Wester

EM No 28. NOTES FROM THE FIFTH INTERNATIONAL SWESAT CONFERENCE. Umeå, May 31 – June 2, 1997. Christina Stage

1998

EM No 29. A COMPARISON BETWEEN ITEM ANALYSIS BASED ON ITEM RESPONSE THEORY AND ON CLASSICAL TEST THEORY. A Study of the SweSAT Subtest WORD. Christina Stage

EM No 30. A COMPARISON BETWEEN ITEM ANALYSIS BASED ON ITEM RESPONSE THEORY AND ON CLASSICAL TEST THEORY. A Study of the SweSAT Subtest ERC. Christina Stage

EM No 31. NOTES FROM THE SIXTH INTERNATIONAL SWESAT CONFERENCE. San Diego, April 12, 1998. Christina Stage

1999

EM No 32. NONEQUIVALENT GROUPS IRT OBSERVED SCORE EQUATING. Its Applicability and Appropriateness for the Swedish Scholastic Aptitude Test. Wilco H.M. Emons

EM No 33. A COMPARISON BETWEEN ITEM ANALYSIS BASED ON ITEM RESPONSE THEORY AND ON CLASSICAL TEST THEORY. A Study of the SweSAT Subtest READ. Christina Stage

EM No 34. PREDICTING GENDER DIFFERENCES IN WORD ITEMS. A Comparison of Item Response Theory and Classical Test Theory. Christina Stage

EM No 35. NOTES FROM THE SEVENTH INTERNATIONAL SWESAT CONFERENCE. Umeå, June 3–5, 1999. Christina Stage

2000

EM No 36. TRENDS IN ASSESSMENT. Notes from the First International SweMaS Symposium Umeå, May 17, 2000. Jan-Olof Lindström (Ed)

EM No 37. NOTES FROM THE EIGHTH INTERNATIONAL SWESAT CONFERENCE. New Orleans, April 7, 2000. Christina Stage

2001

EM No 38. NOTES FROM THE SECOND INTERNATIONAL SWEMAS CONFERENCE, Umeå, May 15-16, 2001. Jan-Olof Lindström (Ed)

EM No 39. PERFORMANCE AND AUTHENTIC ASSESSMENT, REALISTIC AND REAL LIFE TASKS: A Conceptual Analysis of the Literature. Torulf Palm

EM No 40. NOTES FROM THE NINTH INTERNATIONAL SWESAT CONFERENCE. Umeå, June 4–6, 2001. Christina Stage

2002

EM No 41. THE EFFECTS OF REPEATED TEST TAKING IN RELATION TO THE TEST TAKER AND THE RULES FOR SELECTION TO HIGHER EDUCATION IN SWEDEN. Widar Henriksson, Birgitta Törnkvist

2003

EM No 42. CLASSICAL TEST THEORY OR ITEM RESPONSE THEORY: The Swedish Experience. Christina Stage

EM No 43. THE SWEDISH NATIONAL COURSE TESTS IN MATHEMATICS. Jan-Olof Lindström

EM No 44. CURRICULUM, DRIVER EDUCATION AND DRIVER TESTING. A comparative study of the driver education systems in some European countries. Henrik Jonsson, Anna Sundström, Widar Henriksson

2004

EM No 45. THE SWEDISH DRIVING-LICENSE TEST. A Summary of Studies from the Department of Educational Measurement, Umeå University. Widar Henriksson, Anna Sundström, Marie Wiberg

EM No 46. SweSAT REPEAT. Birgitta Törnkvist, Widar Henriksson

EM No 47. REPEATED TEST TAKING. Differences between social groups. Birgitta Törnkvist, Widar Henriksson

EM No 49. THE SWEDISH SCHOLASTIC ASSESSMENT TEST (SweSAT). Development, Results and Experiences. Christina Stage, Gunilla Ögren

EM No 50. CLASSICAL TEST THEORY VS. ITEM RESPONSE THEORY. An evaluation of the theory test in the Swedish driving-license test. Marie Wiberg

EM No 51. ENTRANCE TO HIGHER EDUCATION IN SWEDEN. Christina Stage

Em No 52. NOTES FROM THE TENTH INTERNATIONAL SWESAT CONFERENCE. Umeå, June 1–3, 2004. Christina Stage

2005

Em No 53. VALIDATION OF THE SWEDISH UNIVERSITY ENTRANCE SYSTEM. Selected results from the VALUTA-project 2001–2004. Kent Löfgren

Em No 54. SELF-ASSESSMENT OF KNOWLEDGE AND ABILITIES. A Litterature Study. Anna Sundström

2006

Em No 55. BELIEFS ABOUT PERCEIVED COMPETENCE. A literature review. Anna Sundström

Em No 56. VALIDITY ISSUES CONCERNING REPEATED TEST TAKING OF THE SWESAT. Birgitta Törnkvinst, Widar Henriksson

Em No 57. ECTS AND ASSESSMENT IN HIGHER EDUCATION. Conference Proceedings. Kent Löfgren

Em No 58. NOTES FROM THE ELEVENTH INTERNATIONAL SweSAT CONFERENCE. Umeå, June 12–14, 2006. Christina Stage

2007

Em No 59. PROCEEDINGS FROM THE CONFERENCE: THE GDE-MODEL AS A GUIDE IN DRIVER TRAINING AND TESTING. Umeå, May 7–8, 2007. Widar Henriksson, Tova Stenlund, Anna Sundström, Marie Wiberg