Differential item functioning : types of DIF and observed score based detection methods
Hessen, D.J.

Citation for published version (APA):
Hessen, D. J. (2003). Differential item functioning : types of DIF and observed score based detection methods

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
1 Introduction

Psychological tests are used in a wide range of applied and scientific research settings to measure differences in behavior, cognitive abilities, achievements, personality traits, attitudes, opinions, and other characteristics between individuals or groups or within the same individual or group under different circumstances. For example, clinical uses of tests include the detection of intellectual deficiencies or the examination of persons with emotional disorders or behavioral problems. Examples of educational uses of tests are the classification of children with reference to their ability, the identification of slow and fast learners, educational and occupational counseling, and the selection of students for secondary education. Another example of applied psychological testing is to be found in the selection and classification of industrial personnel or civil servants. In situations as hiring, job assignment, transfer, promotion, or termination, many psychological tests have been proved helpful. In addition to applied psychological test uses, psychological tests are also frequently used in basic research as a means of gathering data for testing scientific hypotheses about, for example, the organization of psychological traits, group differences, and biological and cultural factors associated with specific behavior.

The most important aspect of a psychological test is its validity. The validity of a test is the extent to which the test meets its purpose or measures what it was designed to measure. The validity of a test can be affected by both unsystematic and systematic errors of measurement. In general, the term 'bias' is used to refer to the systematic inaccuracy of measurement. When the purpose of measurement is to study between-group differences on a psychological variable, then the test used ought to be free of measurement bias with respect to group membership. In other words, a test should measure the same unobservable psychological variable in the same way for
different groups under study to be valid for the purpose of between-group comparisons. For example, when the purpose of measurement is the comparison of different groups on arithmetic ability, then observed mean test score differences between subgroups should indicate true mean differences in arithmetic ability. An observed mean test score difference between two groups should not reflect a mean difference in something else the test is measuring unintendedly in addition to arithmetic ability, such as reading ability.

A sufficient condition for a test to be free of group-level measurement bias is that each individual item of the test is free of bias with respect to group membership. In this thesis, the interest is in this type of group-level item bias, for which the more commonly accepted phrase ‘Differential Item Functioning’ (DIF) is used.

Group-level item bias can be distinguished from individual-level item bias. Group-level item bias implies individual-level item bias, but individual-level item bias does not imply group-level item bias. If an item does not measure the same psychological variable in the same way for two groups, then it does not measure the same psychological variable in the same way for all individuals. However, if it does measure the same psychological variable in the same way for all individuals, then the item is also valid for measuring group differences. Conversely, if an item is not appropriate for comparisons between individuals, it still can be valid for between-group comparisons. For example, suppose that an item is designed to measure arithmetic ability, and that persons with a higher reading ability level have on average higher item scores than persons with the same arithmetic ability level but with a lower reading ability level. In this case, the item also measures reading ability at the individual level, and so the item can not be used to compare individuals on just arithmetic ability. However, if men and women do not actually differ in reading ability, the item still can be used to compare these two groups on arithmetic ability.

Since items that function differentially at the group-level decrease the validity of a test for both between-group comparisons and the comparison of individuals, they are undesirable in the test and should be detected and omitted. After test data have been collected, items can be checked for DIF,
for which a variety of statistical methods have been proposed (Millsap & Everson, 1993). By means of such statistical methods the empirical data can be thoroughly investigated for DIF, and the data belonging to items that are flagged as DIF items can be omitted from subsequent statistical analyses.

The concept of DIF typically fits within the framework of Item Response Theory (IRT), in which measurement models have been developed at the item level. In an IRT model, observable item scores are related to a not directly measurable psychological variable, which is assumed to underlie performance on the items. Because the psychological variable is not directly observable, it is called latent. Many IRT models can be considered, depending on the nature of the item scores and that of the latent variable. In addition, various functional relationships can be assumed between the item scores and the latent variable. The DIF studies in this thesis are however restricted to IRT models that specify a relationship between ordinal dichotomous item scores and a unidimensional continuous latent trait.

Many psychological tests exist, that consist of dichotomously scored items which are assumed to measure only one and the same latent trait. These models are especially relevant for the measurement of cognitive abilities and achievements by means of items to which either a correct or an incorrect response can be given.

Fundamental to such IRT models for dichotomous item scores is the Item Response Function (IRF), which relates the probability of a correct or positive item score to the latent trait. Through specifying different assumptions about the shape of the IRFs of the items in the test, and about the relationships between the IRFs, different IRT models for ordinal dichotomous item scores can be defined.

A distinction can be made between nonparametric and parametric IRT models for dichotomous item scores. In nonparametric IRT models for dichotomous item scores, the IRFs are subjected to order restrictions. One such a nonparametric IRT model is Mokken's (1971) Monotone Homogeneity (MH) model. Under the MH model, the IRFs are only assumed to be nondecreasing functions of the latent trait. This assumption of nondecreasingness logically follows from the idea that the probability of giving a correct
or positive response must be higher for persons with a higher latent trait value than for persons with a lower latent trait value. In the MH model, no further assumptions than nondecreasingness are imposed on the IRFs, and the IRFs are therefore allowed to intersect. When the additional assumption of Invariant Item Ordering (IIO) is imposed on the model (Sijtsma & Junker, 1996), which means that the IRFs are not allowed to intersect, then another nonparametric IRT model is formulated, which is called the Double Monotonicity model (Mokken, 1971). In contrast to nonparametric IRT models, different parametric IRT models for dichotomous item scores are defined by the choice of a particular class of parametric functions for the IRFs, such as logistic curves or normal ogives. Well-known parametric IRT models for ordinal dichotomous item scores are the one-parameter logistic model (Rasch, 1960), and the two- and three-parameter logistic models (Birnbaum, 1968).

In the following section, the general definition of DIF under unidimensional latent variable models, the definition of DIF under unidimensional IRT models for dichotomous item scores, and some DIF related invariance conditions are discussed. The second section deals with general problems that are encountered in applying statistical DIF detection methods, and with important statistical results that are relevant for DIF investigations in an IRT context. The last section of this introduction gives an overview of the DIF studies in this thesis.

### 1.1 Differential Item Functioning

In general, DIF means that the probability density or distribution function of an item score is not the same for different subpopulations at some levels of the latent variable assumed to underly performance on the item. If the variable $X$ denotes the item score, the variable $G$ denotes group membership, and $\Theta$ denotes the latent variable, then DIF is defined as (Mellenbergh, 1989)

$$f(X = x \mid G = g, \Theta = \theta) \neq f(X = x \mid \Theta = \theta),$$

(1.1)

for at least one $x$, one $g$ and one $\theta$. In Equation 1.1, $f(\cdot)$ is a probability density function in case of a continuous item score, and it is a probability
distribution function in case of a discrete item score. Note that the definition is general in the sense that it is independent of the measurement levels of the three variables \( X, G \) and \( \Theta \).

For an IRT model for dichotomous item scores and a unidimensional latent trait the definition of DIF in Equation 1.1 is reduced to

\[
P(X = 1 \mid G = g, \Theta = \theta) \neq P(X = 1 \mid \Theta = \theta), \tag{1.2}
\]

for at least one \( g \) and one \( \theta \). In Equation 1.2, \( P(\cdot) \) denotes probability, \( X \) is the dichotomous item score with realization 1 for a positive or correct answer to the item (and 0 for a negative or incorrect answer), and the variable \( \Theta \) denotes the unidimensional latent trait. Thus, under an IRT model for dichotomous item scores DIF is defined to exist if the IRFs of the same item are not identical in different subpopulations.

The opposite condition of DIF defined by Equation 1.2, is the condition that the item score \( X \) and group membership \( G \) are statistically independent at all fixed values of \( \Theta \). Formally, this opposite condition is

\[
P(X = 1 \mid G = g, \Theta = \theta) = P(X = 1 \mid \Theta = \theta), \text{ for all } g \text{ and all } \theta. \tag{1.3}
\]

Since \( \Theta \) is not directly observable, the condition in Equation 1.3 is called Unobserved Conditional Invariance (UCI) of a dichotomous item score with respect to group membership (Millsap & Meredith, 1992). Note that Equation 1.3 is not a definition of unidimensionality. The probability of a positive or correct item score can still depend on some other variable than group membership in addition to the latent trait. Therefore, UCI is only a necessary but not a sufficient condition for the item score to be a valid measure of the intended latent trait for a between-group comparison.

The condition

\[
P(X = 1 \mid G = g, Y = y) = P(X = 1 \mid Y = y), \text{ for all } g \text{ and all } y, \tag{1.4}
\]

in which the latent trait \( \Theta \) is substituted by an observable variable \( Y \) that is used as a proxy of the latent trait \( \Theta \), is called Observed Conditional Invariance (OCI) of a dichotomous item score with respect to group membership (Millsap & Meredith, 1992). Millsap and Meredith (1992) pointed
out that UCI and OCI are not in general equivalent, that is, OCI does not imply, and is not implied by UCI. However, they showed that when the following two conditions hold simultaneously, then UCI and OCI are equivalent. First, the observable variable $Y$ must be ‘Bayes sufficient’ for the latent trait $\Theta$ (Lehmann, 1986, sec. 1.9). Bayes sufficiency of $Y$ means that the item score $X$ and the latent trait $\Theta$ are statistically independent at all fixed values of $Y$. Second, $Y$ and $G$ must be statistically independent at all fixed values of $X$ and $\Theta$, which means that the observable variable $Y$ must be independent of group membership given $X$ and $\Theta$. This last condition is equivalent to the condition that the observable variable $Y$ itself is free of DIF or unbiased for group-level measurement at each of the two levels of the dichotomous item score $X$.

1.2 Detection of Differential Item Functioning

In the literature, DIF detection methods are usually described for the simplest situation of two groups ($G = 1, 2$); see Millsap & Everson (1993) for an overview. Traditionally, a distinction is made between a ‘reference’ group (often assumed to be the majority group in the population) and a ‘focal’ group (the minority group). This distinction is however quite arbitrary when the two groups under study do not differ in population or sample size and, theoretically, the distinction is of no great importance, because it has no implications for most statistical DIF detection methods.

If the dichotomous data of two groups are assumed to follow an IRT model, like one of the nonparametric or parametric models mentioned in the introductory section, and if item $k$ is the item studied for DIF, then the null hypothesis to be tested by a statistical DIF detection method should be

$$H_0: P(X_k = 1 | G = 1, \Theta = \theta) = P(X_k = 1 | G = 2, \Theta = \theta), \text{ for all } \theta. \quad (1.5)$$

This null hypothesis is equivalent to UCI of the dichotomous item $k$ in the two-group situation. If there is no a priori knowledge about a departure from this null hypothesis, the null hypothesis should be tested against the
broad alternative hypothesis

\[ H_1: P(X_k = 1 \mid G = 1, \Theta = \theta) \neq P(X_k = 1 \mid G = 2, \Theta = \theta). \quad (1.6) \]

for at least one \( \theta \). This alternative hypothesis is equivalent to the definition of DIF under an IRT model for dichotomous item scores in the two-group situation. However, if there is a priori knowledge about a specific departure from the null hypothesis, then a more specific alternative hypothesis can be used in order to enhance the power of the statistical test. Specific alternative hypotheses can be formulated on the basis of the different types of DIF that can exist for an IRT model that fits the data of each of the two groups separately (see Chapter 2).

From parametric IRT models for dichotomous item scores DIF detection methods follow which operate within the measurement model, and are therefore based on an UCI model. These DIF detection methods are procedures for testing the equality of item parameter estimates between groups. In general, these procedures require large sample sizes, which could be a problem in practical situations. Moreover, parametric IRT models may be too restrictive and thus may not fit the empirical data adequately. Consequently, the tests of DIF that are incorporated in a parametric measurement model can be confounded with tests of fit.

Alternative methods for DIF detection are methods in which an observed score is used as a proxy for the latent trait. If \( Y \) denotes the observed score on which the examinees of two groups are matched, then the statistical DIF detection method is based on the null hypothesis

\[ H_0: P(X_k = 1 \mid G = 1, Y = y) = P(X_k = 1 \mid G = 2, Y = y), \text{ for all } y. \quad (1.7) \]

This null hypothesis is equivalent to OCI of the dichotomous item \( k \) in the two-group situation.

As is already mentioned in Section 1.1, DIF detection methods based on an OCI model only test for UCI when the observed matching variable \( Y \) itself does not function differentially with respect to group membership and when it is Bayes sufficient for the latent trait, that is, when the observed matching variable \( Y \) captures all relevant information in the latent trait for
the studied item $k$. When the observed matching variable $Y$ is a sufficient statistic for the latent trait, or in case of perfect reliability when there is a one-to-one correspondence between $Y$ and the latent trait, then Bayes sufficiency of $Y$ for the latent trait holds.

A problem in the application of DIF detection methods based on an OCI model is the choice of the observed matching score $Y$, which must be Bayes sufficient and free of DIF. In case the items of a test are assumed to be measures of the same latent trait, usually a function of the item scores is selected as the matching variable. The most obvious choice of $Y$ is the unweighted sum of all test item scores without the item score of the studied item $k$. This sum score is usually called the rest score. The rest score is the most obvious choice of the matching variable because it is directly observable, it can be computed easily, and because it does not include the studied item $k$ that should paradoxically be free of DIF to be legitimately included in the matching variable. If the data of the items included in the matching variable follow the rather restrictive one-parameter logistic model (Rasch, 1960), then none of these items functions differentially and the unweighted sum of the corresponding item scores is a sufficient statistic for the latent trait (Fischer, 1995) and, therefore, also Bayes sufficient. So if the Rasch model fits the empirical data of the rest score items, then a DIF detection method based on an OCI model with the rest score as matching variable can be diagnostic of UCI and DIF for the studied item $k$ that is not included in the matching variable. However, if the data of the item scores included in the matching variable do not follow the one-parameter logistic Rasch model, but a less restrictive two- or three-parameter logistic model (Birnbaum, 1968) or one of Mokken’s nonparametric IRT models, then the unweighted sum over any subset of these item scores is unfortunately neither a sufficient statistic, nor Bayes sufficient for the latent trait.

Two well-known statistical techniques that can be used for the detection of DIF and that are both based on an OCI model are the Mantel-Haenszel procedure (Mantel & Haenszel, 1959; Holland & Thayer, 1988) and the Logistic Regression procedure (Swaminathan & Rogers, 1990a). In both procedures, the null hypothesis to be tested is based on the null hypothesis in Equation 1.7, and therefore, both procedures are not diagnostic of UCI
or DIF in general.

The Mantel-Haenszel procedure is designed to test its null hypothesis against an alternative hypothesis of only one specific departure or type of DIF, whereas the Logistic Regression procedure is designed to test its null hypothesis against the simultaneous alternative hypothesis of two specific departures or types of DIF (the null and alternative hypotheses of both procedures and the different types of DIF will be explained and defined in later chapters).

In both the Mantel-Haenszel and the Logistic Regression procedure, it is common practice to use the total score (the unweighted sum of all test item scores) as the matching variable. Using the total score with the inclusion of the item score of a studied item \( k \) that actually functions differentially, seems to contradict the condition that the matching variable must be free of DIF. However, Holland and Thayer (1988) showed that under the one-parameter logistic Rasch model the inclusion of the item score of the studied item \( k \) in the matching variable is a necessary condition for the Mantel-Haenszel null hypothesis to be equal to the IRT condition of UCI with respect to group membership, or equivalently, to the IRT null hypothesis of no DIF.

### 1.3 Overview of the thesis

Since the following chapters have initially been written as research articles, there is some overlap in the introductory sections. Each of the chapters deals with one or more types of DIF, which are all defined formally in the second chapter. In the last four chapters, the focus is on the utility of two observed score based methods in detecting one or more of these types of DIF. Since the chapters are independent of one another, each chapter can be read without first reading one of the preceding chapters.

In the second chapter, the concept of DIF under a general nonparametric IRT model for dichotomous item scores is studied theoretically. The concept of DIF is related to the concept of unidimensionality. Furthermore, four DIF distinctions are discussed, that each define two complementary types of DIF. Throughout the chapter, two types of DIF are called com-
plementary, when they are mutually exclusive (disjoint) and exhaustive. The interrelationships between types of DIF that are not complementary by definition, are examined, and it is shown that given the general non-parametric IRT model and the four DIF distinctions only seven mutually exclusive types of DIF can exist. Moreover, it is studied which of these types of DIF are excluded or can exist for several parametric and nonparametric IRT models. Each type of DIF that can exist for a specific model can be the basis for the alternative hypothesis in a corresponding DIF detection method under that model.

In the third chapter, the robustness and the power of the Logistic Regression (LR) procedure in detecting the two types of DIF it is designed to detect, are investigated. The dichotomous data are generated under Mokken’s Monotone Homogeneity (MH) model in order to study the utility of the LR procedure under a more general model than the one-, two- or three-parameter logistic models.

In the fourth chapter, the effect of the inclusion of the studied item in the matching variable and the effect of using the correction for continuity in the chi-square test statistic on both the robustness and the power of the Mantel-Haenszel procedure are studied. The dichotomous data are simulated under the two-parameter logistic model to explore the use of the Mantel-Haenszel procedure under a violation of the necessary condition for the rest score and the total score to be Bayes sufficient for the latent trait.

In the fifth chapter, the Mantel-Haenszel procedure is subjected to a second simulation study. In this study, the focus is on the effect of item discrimination on the robustness and the power of the Mantel-Haenszel procedure. Again, data are generated under the two-parameter logistic model, but now mainly to evaluate the effect of item discrimination on the Mean Squared Error (MSE) and bias of the Mantel-Haenszel odds-ratio estimator, that is used as a DIF effect size measure.

Finally, in the sixth chapter, a new class of so-called Constant Latent Odds-Ratios (CLORs) models is introduced, under which it is shown that the total score is a sufficient statistic for the latent trait. The chapter also deals with a new class of parametric special cases of the most general CLORs model, in order to show that this general CLORs model is not equiv-
alent to the one-parameter logistic model. Next, it is shown that methods based on an OCI model in general and the Mantel-Haenszel procedure in particular can be diagnostic of UCI or DIF under the whole class of CLORs models. Furthermore, hypothetical situations are discussed that can arise in practice when the Mantel-Haenszel procedure is used for DIF detection under the condition that one of the CLORs models fits the empirical data.