Edukacja 2014, 5(130), 92–111
An interdisciplinary approach
ISSN 0239-6858

Comparison of Mantel–Haenszel with IRT procedures for DIF detection and effect size estimation for dichotomous items

Bartosz Kondratek, Magdalena Grudniewska

Student Performance Analysis Unit, Educational Research Institute*

The article compares two methods used to detect differential item functioning (DIF) of dichotomously scored items: a nonparametric solution based on the Mantel–Haenszel procedure (MH) and a parametric IRT approach with a likelihood ratio test. A Monte Carlo experiment was performed in order to evaluate performance of both statistics in various conditions of DIF uniformity. Results confirmed the theoretical prediction that the MH test has greater statistical power in detecting uniform DIF than the likelihood ratio test and less power than the LR test in cases of non-uniform DIF. Apart of examining statistical power of the test, specific measures of DIF effect size were compared: *MH D–DIF* and three measures of *P–DIF* expressed on the item easiness scale.

Keywords: differential item functioning, Mantel–Haenszel test, item response theory.

Differential item functioning (DIF) is a statistical term that describes a condition when the item response is dependent not only on the ability level of the subject but also on the value of some additional group membership variable. If DIF is present verification plays an important role in psychometric analysis of a test and is closely related to the problem of validity.

Let U_i denote response to item i, θ be the level of ability that the test measures and G be the group membership variable, then the general equation defining DIF with regard to group membership will be of the form (c.f. Penfield and Camilli, 2007):

$$U_i|\theta, G \neq U_i|\theta,$$

which states that conditional distribution of the response is not explained solely by the ability variable (θ) of the examinee but is additionally related to which group (G) the examinee belongs to. In the case of dichotomously scored item it can be rewritten as:

$$\mathrm{P}(\mathrm{U_i} = 1|\theta, \mathrm{G}) \neq \mathrm{P}(\mathrm{U_i} = 1|\theta),$$

which means that the probability of correct response to the item is dependent not only on the ability θ, but also on the group membership G. If G is two valued, $G \in \{f, r\}$, then differential item functioning of item i can be also expressed as:

$$P(U_i = 1|\theta, G = f) \neq P(U_i = 1|\theta, G = r), \quad (1)$$

meaning that the probability of correct response of an examinee with ability θ in group

Article based on research carried out within the systemic project "Quality and effectiveness of education – strengthening of institutional research capabilities" executed by the Educational Research Institute and co-financed from the European Social Fund (Human Capital Operational Programme 2007–2013, Priority III High quality of the education system). A preliminary version of this article was published primarily in Polish in *Edukacja*, *122*(2) 2013.

* Address: Zespół Analiz Osiągnięć Uczniów, Instytut Badań Edukacyjnych, ul. Górczewska 8, 01-180 Warszawa, Poland. Email: b.kondratek@ibe.edu.pl

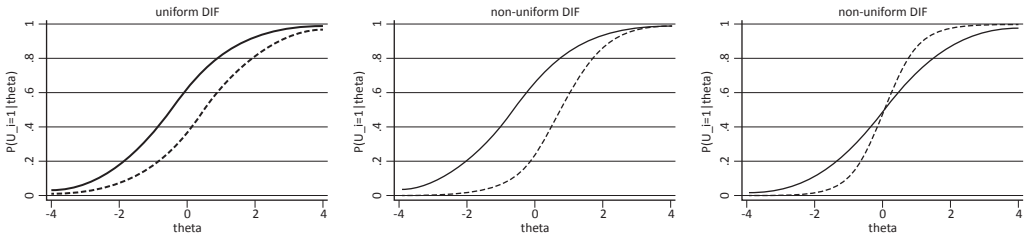

Figure 1. Examples of DIF (continuous line depicts $P(U_i = 1|\theta)$ for: $G = r$, dashed line for: $G = f$).

f differs from the probability of correct response of an examinee with the same level of ability in group *r*.

Figure 1 collects three examples of differential item functioning as defined by eq. (1) by means of curves that depict the probability of correct response conditional on θ (item characteristic curve, ICC). The left-hand graph shows the uniform DIF – the ICC in one group is shifted parallel to the ICC for the same item in the other group. All other cases of DIF will be a non-uniform DIF. In the middle graph the item *i* is easer in group *r* at all levels of θ, just like in the leftmost graph, however the magnitude of discrepancy in difficulty conditional on θ is different at different levels of θ. The rightmost graph presents an interesting case of non-uniform DIF, namely, for examinees of ability $\theta < 0$ the item is easier for group *r*, however for examinees of ability $\theta > 0$ the same item is easier for group *f*.

Pioneering work regarding DIF analysis dates back to the 1960s in the USA, when the need to identify test items biased with regard to minorities became acute. Hence in DIF analysis a classical unsymmetrical division into two groups is present: the focal minority group and the reference majority group. In this article group membership indexes *f* and *r* will be used in order to comply with this tradition however it is worth noting that DIF analysis in educational studies is often performed when the grouping variable divides examinees more evenly and without any obvious indication of which group is more likely to be measured unfairly by the test – gender is probably the best example.

The term item bias refers to the situation when one group is being favoured over another group as a consequence of item content unrelated to the ability intended to be measured by the test. Item bias is a specific distortion of validity of the test and is not equivalent to the presence of DIF. Differential item functioning states that the item response is related to some additional factor that at the same time is unrelated to the ability measured by the test as a whole but is correlated with group membership. DIF is a necessary condition for item bias, but is not a sufficient condition for item bias. Flagging an item as biased requires an expert analysis of the item content in the context of possible causes of DIF. It is possible, that the item-specific factor causing DIF is actually an important part of test's content domain which is not represented in other items of the test, thus inclusion of such an item presents no hazard to validity and will not discriminate against any of the groups (see Zieky, 1993).

It is also worth mentioning the difference between DIF and between-group differences in ability. The very essence of the concept of DIF is to distinguish the actual differences in the ability level between groups from differences in the way the item behaves because of factors other than the ability measured by the test as a whole. Conditioning over θ, which is present in the definition of DIF, indicates that the analysis is performed under control of differences in distribution of ability between groups.

According to what was stated above it can be concluded that DIF detection for dichotomously scored items will require analysis of

item difficulty conditional on group membership G with statistical control over the ability variable θ. Operationally the level of ability is usually defined within the test as some form of score obtained from the whole test. The first and natural solution for the problem stated above was to employ the Mantel–Haenszel (MH) test. Very popular in the setting of clinical trial data analysis, the MH test allows analysis of statistical significance of differences in the distribution of a dichotomous outcome variable between two groups stratified on some additional variable that is significantly related to the outcome. The MH test is also called Cochran––Mantel–Haenszel test, to credit Cochran who proposed a very similar procedure earlier (Agresti, 2002). An alternative approach to DIF analysis, to be covered in the article arose with the rapid development of item response theory (IRT) in the last decades of the 20th century. In IRT the relationship between item response and the ability level is modelled explicitly.

The article begins with a brief introduction to both methods of DIF analysis together with specific DIF effect-size measures that can be constructed in each of the approaches. Determining the actual magnitude of DIF is of no less practical importance than significance analysis, hence the effect-size topic will organise the logic of presentation of the methods. Afterwards, a Monte Carlo experiment comparing the performance of both methods under various conditions is presented.

Mantel–Haenszel test DIF analysis

In the MH approach responses to a dichotomous item analysed for DIF between two groups are stratified on the number of sum score points obtained in the test. A contingency table of size $2 \times 2 \times M$ is thus obtained, where M is the total number of sum score points. Probabilities of observing a particular response to the item conditional on group membership and sum score category m are denoted in the following manner:

Group	Item response		Total
	1	0	
f	p_{1fm}	p_{0fm}	p_{fm}
r	p_{1rm}	p_{0rm}	p_{rm}
Total	p_{1m}	p_{0m}	p_m

The MH test is constructed in terms of odds ratios. The odds of a correct response are the probability of correct response divided by the probability of incorrect response. Hence the ratio of such odds for examinees from groups r and f within score category m is given as:

$$\alpha_m = \frac{p_{1rm}/p_{0rm}}{p_{1fm}/p_{0fm}}$$

Following the above designations the null hypothesis and the alternative hypothesis of MH test can be stated as (c.f. Dorans and Holland, 1993):

H_0: $\alpha_m = 1$ $\quad m \in \{1, \ldots, M\}$,
H_1: $\alpha_m = \alpha \neq 1$ $\quad m \in \{1, \ldots, M\}$.

The null hypothesis states that the odds of correct response to the item are the same in both groups in every score category m. It can be equivalently rewritten in the manner that DIF was defined in eq. (1):

$$H_0: P(U_i = 1|m, f) = P(U_i = 1|m, r), m \in \{1, \ldots, M\}$$

This means that probability of correct response to the item is not related to group membership as long as the test score m is taken into account. What is unique to the MH test is the way the alternative hypothesis is stated. The H_1 of the MH test states that the difference of these conditional-on-the-score category probabilities will be nonzero in a constant direction. Moreover, according to H_1 all the odds ratios α_m will equal one common odds ratio α. The number of observations can be indexed analogously as the earlier probabilities:

Group	Item response		Total
	0	1	
r	N_{0rm}	N_{1rm}	N_{rm}
f	N_{0fm}	N_{1fm}	N_{fm}
Total	N_{0m}	N_{1m}	N_m

The MH statistic with correction for continuity can be expressed as:

$$MH_{\chi^2} = \frac{\left(\sum_{m=1}^{M}\left(N_{1fm} - E(N_{1fm})\right) - 0{,}5\right)^2}{\sum_{m=1}^{M} D^2(N_{1fm})}, \quad (2)$$

where $E(N_{1fm})$ and $D^2(N_{1fm})$ are the expected value and the variance of observations N_{1fm} under H_0. Under H_0 the MH_{χ}^2 statistic is asymptotically χ^2 distributed with one degree of freedom (Dorans and Holland, 1993).

It was proven (Radhakrishna, 1965) that the MH test is a uniformly most powerful test of a null hypothesis of conditional independence of proportions between groups, if the hypothesis of constant odds ratio is valid. If the hypothesis of constant odds ratio is not valid the MH test reduces in power. It means that the MH test will perform less well in detecting non-uniform DIF in comparison to procedures that allow for interaction between DIF magnitude (defined as odds ratio) and the ability level (Swaminathan and Rogers, 1990). From eq. (2) it can be deducted that in the extreme situation when α_m are related to m in such a way, that some of α_m are above 1 and some α_m are below 1, the respective discrepancies of N_{1fm} from their expected values will cancel out. Owing to such dependence of performance of the MH test on the assumption of constant odds ratio, some additional procedures testing for violation of this hypothesis are often performed, i.e. the Wolf test (1955).

Mantel and Haenszel (1959) also proposed an estimator of the common odds ratio of the form:

$$\alpha_{MH} = \frac{\sum_{m=1}^{M} p_{1rm} p_{0fm} N_m}{\sum_{m=1}^{M} p_{1fm}\, p_{0fm} N_m}, \quad (3)$$

in which more weight is applied to cells with higher marginal totals N_m. For an item that is easier (conditioned on ability) for group r we would obtain $\alpha_{MH} > 1$ and $\alpha_{MH} < 1$ in the opposite case.

IRT likelihood ratio test DIF analysis

Discussion of DIF analysis in IRT will be limited in this article to the case of the two-parameter logistic model (2PLM), however the ideas presented could be easily extended to other models, in particular in the case of politomously scored items. Thissen, Steinberg and Wainer (1993) provided a general presentation of DIF testing within IRT modelling (to learn about other methods of DIF analysis see Penfield and Camilli, 2007).

In IRT the relationship between the probability of observing a correct response to the item n and the level of examinee's ability θ, that appears in definition of DIF (1), is modelled explicitly. In 2PLM such probability is given by a logistic function that depends on two item parameters b_n and a_n:

$$p_n(\theta) = P(U_n = 1|\theta, a_n, b_n) = \frac{1}{1+e^{-a_n(\theta - b_n)}}. \quad (4)$$

Parameter b_n (item difficulty) shifts the logistic function along θ scale and parameter a_n (item discrimination) is responsible for the steepness of the function at the point of $\theta = b_n$. These two item parameters make 2PLM sensitive to the cases of both uniform and non-uniform DIF (Figure 1).

The full IRT model describes the probability of observing the whole vector of test item responses $U = (U_1, \ldots, U_n, \ldots, U_N)$, not only of the item being analysed for DIF. To abbreviate the notation, assume that $p_n(\theta)$ stands for ICC of item n and all items have ICC of the form of (4) with parameters (a_n, b_n) and that $\psi_G(\theta)$ denote the ability distribution in group $G \in \{f, r\}$. A situation when no DIF is present will be described by an IRT model in which probability of observing particular $U = u$ response vector is given by:

$$P(U = u|G) = \int \left[\prod_{n \in \{1,\ldots,N\}} p_n(\theta)^{u_n} \left(1 - p_n(\theta)\right)^{1-u_n} \right] \psi_G(\theta) d\theta. \quad (5)$$

The product in the square brackets (under condition of local independence of items[1]) is a conditional likelihood function that describes the probability of observing the response vector $U = u$ conditional on level of ability θ and on the parameters that characterise functions p_n. As one can observe, the conditional likelihood does not depend on the group membership, the only group-varied element of the model (5) is ability distribution, ψ_G, over which the conditional likelihood is integrated.

In eq. (5) it is assumed that item parameters for all items are the same in both groups. A situation when there exists DIF for one item is constructed by introduction of a different sets of parameters for this item for examinees from groups *f* and *r* – (a_i^f, b_i^f) and (a_i^r, b_i^r) respectively. A model that is allowing for DIF for item *i* will, therefore, be of the following form:

$$P(U = u|G) = \\ = \int \left[\prod_{n\in\{1,\dots,N\}\setminus\{i\}} p_n(\theta)^{u_n}\left(1 - p_n(\theta)\right)^{1-u_n} \right] \cdot \\ \cdot p_i^G(\theta)^{u_i}\left(1 - p_i^G(\theta)\right)^{1-u_i} \psi_G(\theta) d\theta \quad (6)$$

In this framework the null hypothesis and the alternative hypothesis that it is tested against can be stated as:

$$H_0: a_i^f = a_i^r \wedge b_i^f = b_i^r$$
$$H_1: a_i^f \neq a_i^r \vee b_i^f \neq b_i^r$$

The null hypothesis is verified by likelihood ratio test (LR test), by using the fact that model (5) is nested within model (6). The test statistic has the form:

$$LR = -2ln\left(\frac{L_0}{L_1}\right), \quad (7)$$

where L_0 is the likelihood function computed on the basis of estimates of model (5) and L_1 is the analogous likelihood function for model (6). Degrees of freedom for the *LR* statistic is equal to the difference between number of parameters being estimated in the two models, which is 2 in the case under consideration (one additional difficulty parameter and one additional discrimination parameter).

What may be noticed is that in order to test for DIF within IRT modelling framework, software that estimates a different ability distribution for the focal and the reference group is needed. An IRT model without the multi-group feature would not properly separate the differences in behaviour of item between groups from the differences in ability distribution between groups, which, is the essence of DIF analysis.

Measures of DIF effect size and item DIF classification

The common odds ratio of the MH statistic (3) is a measure of DIF effect size which is difficult to interpret. In order to facilitate the interpretation the value of α_{MH} is transformed in various ways. One of these transformations is $MH\ D-DIF$ obtained in the following manner:

$$MH\ D-DIF = -2.35\ln[\alpha_{MH}]. \quad (8)$$

Such transformation of α_{MH} produces an estimator of DIF effect size with symmetrical distribution, values ranging from -∞ to +∞. $MH\ D-DIF$ has value of 0 when no DIF is present.

The Educational Testing Service (ETS) developed a classification system of DIF effect size that is based on the significance of MH_{χ}^2 (standard significance level of α = 0,05 is assumed) and the value of $MH\ D-DIF$ measure. Items are divided into three disjoint categories: A, B and C (Dorans and Holland,

[1] The assumption of local (conditional) independence of item responses states that when the value of ability parameter θ is known item responses become statistically independent. This assumption is of profound significance not only in the technical context of parameter estimation by maximum likelihood methods but has an important theoretical interpretation. Namely, the notion that level of ability explains all observable interdependence between items mean that the test is unidimensional (Lord and Novick, 1968).

1993; Zieky, 2003) according to the following rules:

- Category A – if MH test result was negative or if MH test result was positive, but the absolute value of *MH D – DIF* is less than 1;
- Category B – if MH test result was positive and absolute value of *MH D – DIF* is between 1 and 1.5 or if MH test was positive and 95% confidence interval around *MH D – DIF* is disjunctive with interval [-1;1];
- Category C – if 95% confidence interval around *MH D – DIF* is disjunctive with interval [-1;1] and absolute value of *MH D – DIF* is above 1.5 (in particular this means positive MH test result).

The above rules are shown in form of a decision tree in Figure 2 which also includes values of α_{MH} that correspond to *MH D – DIF*. Inclusion of α_{MH} in the graph is motivated by the fact, that most statistical packages report the results of MH test on the scale of "raw" odds ratio α_{MH}.

Items in category C require special attention for the test developer with regard to potential bias. Information on the category of given item is supplemented with information about whether the item is more difficult for the focal group (items marked with "-") or more difficult for the reference group (items marked with "+").

MH D – DIF transforms α_{MH} into a more symmetric distribution and facilitates constructing rules for item DIF effect size classification, however it still does not provide a clear interpretation of actual size of the DIF effect. What seems to be a natural choice for measuring DIF effect is to perform it on the scale if item easiness. The question is of how much easier (or more difficult) the item *i* is for group *f* would it function in group *f* the same way as it does in group *r*. A family of DIF effect size measures that aim at answering such a question are suffixed *P – DIF* in this article.

In order to have some insight into how does the ETS DIF categorisation relate to the magnitude of difference in item easiness between groups when ability is controlled, let us note that for each score category *m*, the probability of observing the correct response p_{1rm} can be expressed in terms of α_m and p_{1fm}:

$$p_{1rm} = \frac{\alpha_m p_{1fm}}{1 - p_{1fm} + \alpha_m}.$$

On condition of the validity of the assumption of common odds ratio α_m we can estimate the probability that examinees from group *f* respond correctly to the analysed item in a hypothetical situation that this item functions in group *f* the way it does in group *r*:

$$rp_{1r}^{\dagger} = \frac{\alpha_{MH} p_{1f}}{1 - p_{1f} + \alpha_{MH}}.$$

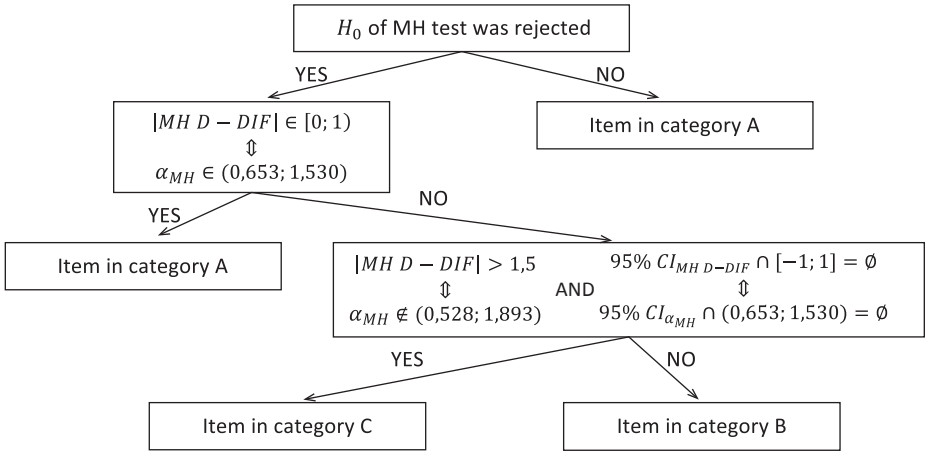

Figure 2. Decision tree for item classification with respect do their DIF based on *MH D – DIF*.

Finally, the pursued difference in easiness of item for group *f* on the basis of MH statistic is (c.f. Dorans and Holland, 1993):

$$MH\ P-DIF = p_{1f} - p_{1r}^{\dagger}. \quad (9)$$

Figure 3 shows how the difference of easiness of an item that results from differential item functioning described by $MH\ P-DIF$ (9) relates to easiness of an item in the focal group p_{1f} and to boundary values of α_{MH} that can be found DIF effect size classification in Figure 2. It can be commented that firstly, boundaries of 95% CI around α_{MH} that define transition between categories A, B and C, depend on how easy the item is in group *f* – items of moderate easiness require larger absolute value of difference of easiness resulted from DIF in order to shift between classification categories than items exhibiting more extreme easiness in group *f*. Secondly the procedure for categorizing DIF size is not symmetric with respect to group membership, group *f* is favoured when $MH\ P-DIF$ is positive and group *r* is favored when $MH\ P-DIF$ is negative. This lack of symmetry is a consequence of adopting symmetric criteria ±1 or ±1.5 around *MH D – DIF* (Figure 2) in order to define classification boundaries where *MH D – DIF* is in fact a nonlinear transformation of α_{MH} (eq. (8)). If we assume that the expected difference of easiness of items between groups due to DIF is an adequate measure of DIF effect size, these two remarks point to a clear drawback of the classification described above. However it should be noted that in the range of easiness p_{1f} between 0.25 and 0.75, in which most items of a well-constructed test should fall, the thresholds for passing between categories A, B and C, are at a similar level to *MH P – DIF*.

Classical test sum scores can be utilised to derive another estimator for difference

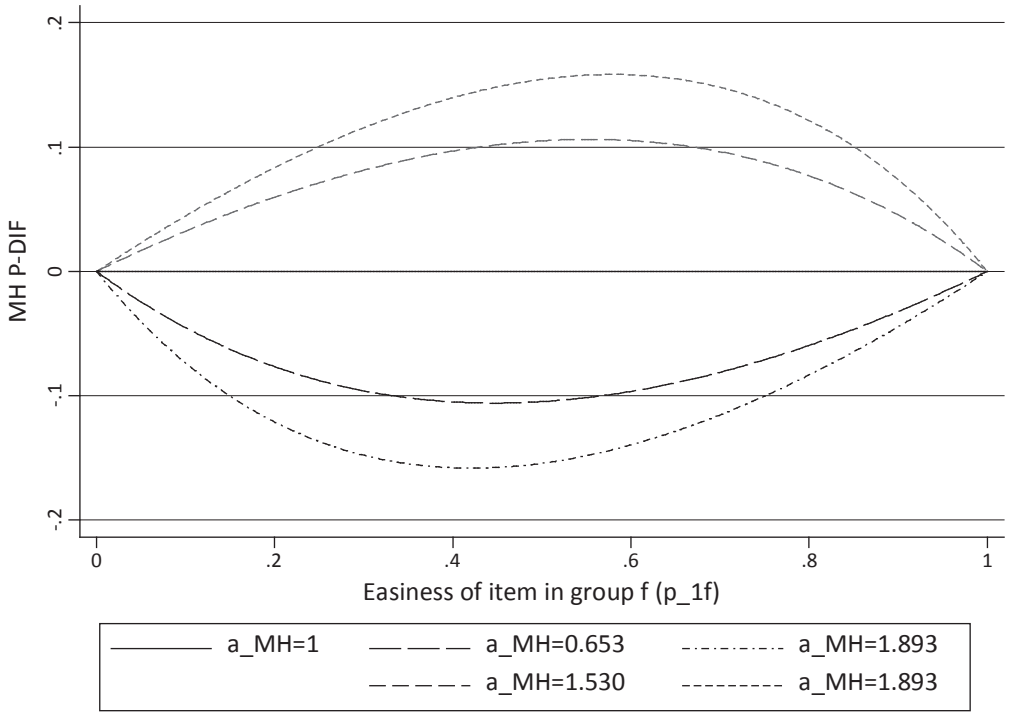

Figure 3. Relation between and the easiness of the item in group *f* for boundary values of α_{MH}.

in easiness of item due to DIF that, in comparison to $MH\ P-DIF$, does not resort to the common odds ratio:

$$STD\ P-DIF=\frac{\sum_{m=1}^{M}N_{fm}\left(p_{1fm}-p_{1rm}\right)}{\sum_{m=1}^{M}N_{fm}}. \quad (10)$$

Analysing the RHS of (10), we see that STD P – DIF is an average of differences between easiness of item between groups *f* and *r* in each score category *m*, weighted by N_{fm} – the number of observations for score category *m* for group *f*. Dorans and Holland (1993) compared $STD\ P-DIF$ and $MH\ P-DIF$ and concluded that while both of them estimated the same parameter, i.e. conditional difference of easinesses, they differed in the manner that cases were weighted. $MH\ P-DIF$ utilises the common odds ratio statistic (3) in which the weights are computed to be optimal in the context of test power. Consequently values computed by $STD\ P-DIF$ and $MH\ P-DIF$ will slightly differ (Dorans and Holland, 1993).

Equation (10) is only a step away from deriving a measure of DIF effect size expressed on the item easiness scale that would incorporate an IRT model. Let p_i^f, designate the item characteristic function of item *i* for group *f* and p_i^r, its counterpart for group *r*, which in case of 2PLM stands for eq. (4) with parameters $(a_i^f,\ b_i^f)$ and $(a_i^r,\ b_i^r)$ respectively. A straightforward measure of DIF effect size for IRT, which we denote $IRT\ P-DIF$ (c.f. Wainer's equation *T*(1), 1993), is:

$$IRT\ P-DIF=\int\left[p_i^f(\theta)-p_i^r(\theta)\right]\psi_f(\theta)d\theta. \quad (11)$$

Equation (11) explicitly expresses the difference between actual easiness of the item *i* in population *f* and the easiness of the same item in *f* would it function according with parameters that characterise it in population *r*. It may be noticed that $STD\ P-DIF$ given by (10) can be interpreted as an nonparametric version of $IRT\ P-DIF$ (10) – in the former integration is performed over discrete sum score divided into categories $m\in\{1,\ldots,M\}$ and in the latter integration is over the continuous latent ability variable θ.

Acknowledging earlier criticism of DIF item classification based on the magnitude of $MH\ D-DIF$ (Figure 2), one can construct an alternative classification utilizing $P-DIF$ effect size measures. Adopting threshold values of Monahan, McHorney, Stump and Perkins (2007):

- Category A – if result of test to verify the hypothesis of no DIF was negative or positive, but with absolute value of $P-DIF$ less than 0.05;
- Category B – if result of test to verify the hypothesis of no DIF was positive and absolute value of $P-DIF$ between 0.05 and 0.1;

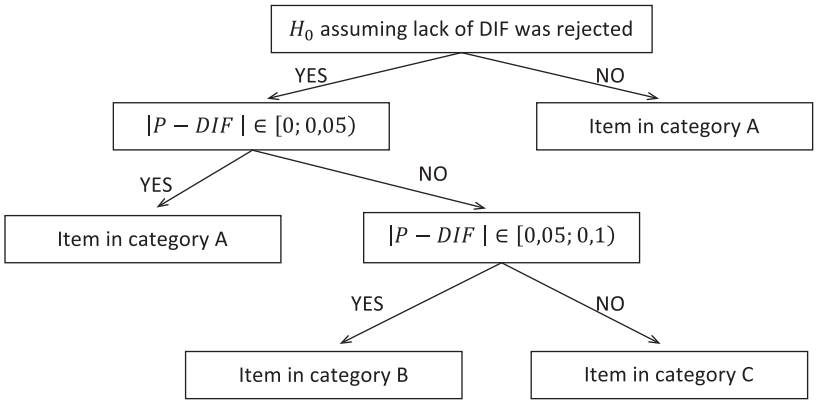

Figure 4. Decision tree for item classification with respect do their DIF based on *P – DIF*.

- Category C – if result of test to verify the hypothesis of no DIF was positive and absolute value $P - DIF$ is above 0.1.

A decision tree for $P - DIF$ governed DIF item classification is shown in Figure 4. The first noticeable feature is that this classification is defined in a general manner, i.e. without referring to any specific test that could be used to test for statistical significance of DIF. Hence, such a classification could be applied either after performing MH test and calculating *MH P – DIF* (9) or *STD P – DIF* (10), or after performing LR test and calculating *IRT P – DIF* (11). The second important feature of this classification is that it does not mention the precision of estimating *P – DIF*, compared to analysis of 95% CI around *MH D – DIF* in the previous classification. Not including information on precision of estimation of *P – DIF* is a disadvantage of this decision tree, which could be easily overcome by inclusion of the magnitude of standard error of a given estimator of *P – DIF* into the procedure. The expression for standard error of *STD P – DIF* estimator can be found in Dorans and Holland (1993), however estimating standard error for the *IRT P – DIF* estimator appears to be more troublesome, potentially requiring a Monte Carlo approach.

Monte Carlo Experiment

In order to compare performance of the MH method for DIF analysis and another based on an IRT model, a Monte Carlo experiment was performed. Data were generated according to an IRT model (6) for a test containing $N = 20$ items and each item had an item characteristic curve belonging to the family of 2PLM (4). The first 19 items of the test had equal parameters in both populations *f* and *r*, i.e. these items did not exhibit any DIF. Difficulty parameters b_n of these no-DIF items were spaced symmetrically around 0 and their values corresponded to the 5^{th}, 10^{th}, …, 95^{th} centiles of the standard normal distribution $N(0;1)$ and discrimination parameters b_n of these items alternately had values 1 and 1.5. This part of the test did not exhibit any DIF, consequence had an information function tuned for optimal measurement of ability of examinees sampled form population $N(0;1)$. Parameters of mentioned 19 items are collected in Table 1.

Distribution of ability in the focal group was standard normal $\psi_f = N(0;1)$. Distribution of ability in the reference group was of the same shape but shifted to the right with the value 0.253, $\psi_r = N(0.253;1)$, which corresponds to a situation when the mean level of ability in group *r* is at the 60^{th} centile

Table 1
Parameters of 19 items not exhibiting DIF used for simulating data

N	b_n	a_n	n	b_n	a_n	n	b_n	a_n
1	-1.65	1	10	0	1.5	11	1.65	1
2	-1.28	1.5				12	1.28	1.5
3	-1.04	1				13	1.04	1
4	-0.84	1.5				14	0.84	1.5
5	-0.68	1				15	0.68	1
6	-0.52	1.5				16	0.52	1.5
7	-0.39	1				17	0.39	1
8	-0.25	1.5				18	0.25	1.5
9	-0.13	1				19	0.13	1

Table 2
Complete set of experimental conditions in the study; "nu" – non-uniform DIF, "u" – uniform DIF, (-) – group f is disadvantaged by the item, (+) – group f is favoured by the item, (0) – non-uniform DIF resulting in null IRT P – DIF effect

IRT P – DIF	Easiness of the 20th item in *f:* 0.5			Easiness of the 20th item in *f:* 0.7		
	$a^r_{20} = 1$	$a^r_{20} = 1.5$	$a^r_{20} = 2$	$a^r_{20} = 1$	$a^r_{20} = 1.5$	$a^r_{20} = 2$
-0.15	nu(-)	u(-)	nu(-)	nu(-)	u(-)	nu(-)
-0.125	nu(-)	u(-)	nu(-)	nu(-)	u(-)	nu(-)
-0.1	nu(-)	u(-)	nu(-)	nu(-)	u(-)	nu(-)
-0.075	nu(-)	u(-)	nu(-)	nu(-)	u(-)	nu(-)
-0.05	nu(-)	u(-)	nu(-)	nu(-)	u(-)	nu(-)
-0.025	nu(-)	u(-)	nu(-)	nu(-)	u(-)	nu(-)
0	nu(0)	no DIF	nu(0)	nu(0)	no DIF	nu(0)
0.025	nu(+)	u(+)	nu(+)	nu(+)	u(+)	nu(+)
0.05	nu(+)	u(+)	nu(+)	nu(+)	u(+)	nu(+)

of ability distribution in group *f*. The task of DIF verification was thus conducted under circumstances of significant difference in mean ability level between groups, in favour of the reference group.

In the Monte Carlo experiment parameters of item indexed with the number 20 were manipulated – this was the item for which performance of the two methods of DIF analysis were compared. In group *f* only two sets of parameters (a^f_{20}, b^f_{20}) were considered:

- $a^f_{20} = 1.5$ and $b^f_{20} = 0$, a condition under which easiness of the 20th item in population *f* equals 0.50;
- $a^f_{20} = 1.5$ and $b^f_{20} = -0.79163$, a condition under which the 20th item is easier than above, with its easiness equal 0.70 in population *f*.

Parameters of the 20th item in the reference group varied to a larger extend so as to allow a vast potential range of DIF conditions when crossed with the two aforementioned cases of parameter values of the 20th item in the focal group. The discrimination parameter a^r_{20} took three different values:

- $a^r_{20} = 1$, a condition of non-uniform DIF due to flatter item characteristic curve in group *r* than in group *f*;
- $a^r_{20} = 1.5$, a condition of uniform DIF;
- $a^r_{20} = 2$, a condition of non-uniform DIF, due to steeper item characteristic curve in group *r* than in group *f*.

Difficulty parameters were chosen in such a manner so as to achieve a set of predefined easiness values with the pair of parameters (a^r_{20}, b^r_{20}) in population *f*, in order to obtain specific "true" values of the *IRT P – DIF* (11) effect. Specifically, a bisection method with Monte Carlo integration was employed in solving the integral (compare to eq. (11)):

$$IRT\ P - DIF = \int \left[p^f_{20}(\theta) - p^r_{20}(\theta)\right] \psi_f(\theta) d\theta$$

with respect to b^r_{20} in order to obtain nine values of $IRT\ P - DIF$ that ranged equally-spaced from -0,150 up to 0,050.

Finally, a set of $2 \times 3 \times 9$ experimental conditions were analysed which are collected in Table 2. Each condition listed in Table 2 was independently replicated 10 000 times and in each replication a set of 1000 examinee response vectors were simulated for each of the groups *f* and *r*. After every replication:

- The MH test verifying DIF for the 20th item was performed with item response being stratified with respect to sum score computed from responses to all 20 items of the test. The estimator $\widehat{\alpha_{MH}}$ and its 95% CI were

computed. STATA's *cc* (case-control) procedure was employed to perform the task.
- The LR test verifying DIF for the 20th item was performed. Fitting of an IRT model without DIF (5) and with DIF (6) was done with MIRT software (Glas, 2010);
- Three $P-DIF$ effect measures were estimated: $MH\widehat{P-DIF}$ (9), $STD\widehat{P-DIF}$ (10) and $IRT\widehat{P-DIF}$ (11);
- DIF classification into three categories according to the two schemes depicted in Figures 2 and 3 was done, in the latter case $IRT\widehat{P-DIF}$ and LR test results were utilised.

The main goals of the research were to:
- compare the sensitivity of MH and LR tests in detecting DIF under different conditions of DIF effect size and DIF type (uniform vs. non-uniform);
- compare the properties (bias and standard error) of three estimators of *P – DIF* effect size (*MH*, *STD*, *IRT*) under the supposition that the true *P – DIF* effect size is the *IRT P – DIF* condition according to which the data were simulated (Table 2);
- compare the two classification schemes of DIF effect size into categories A, B, C.

Results

The primary goal of the experiment was to compare the sensitivity of MH and LR tests in different conditions. Table 3 presents percentage of cases in which the two methods

Table 3
Percentage of replications in which statistically significant result was observed from Mantel–Haenszel and from likelihood ratio test

IRT *P – DIF*	Statistic	Easiness of the 20th item in *f* : 0.5			Easiness of the 20th item in *f*: 0.7		
		$\alpha^r_{20}=1$	$\alpha^r_{20}=1.5$	$\alpha^r_{20}=2$	$\alpha^r_{20}=1$	$\alpha^r_{20}=1.5$	$\alpha^r_{20}=2$
-0.15	MH	100	100	100	100	100	100
	LR	100	100	100	100	100	100
-0.125	MH	100	100	100	100	100	100
	LR	100	100	100	100	100	100
-0.1	MH	99	100	100	100	100	100
	LR	100	100	100	100	100	100
-0.075	MH	91	97	99	95	99	100
	LR	99	94	100	99	96	100
-0.05	MH	54	75	86	60	80	90
	LR	94	62	93	92	68	94
-0.025	MH	14	28	40	14	30	44
	LR	86	19	74	80	21	71
0	MH	7	5	7	7	5	8
	LR	83	5	60	77	5	54
0.025	MH	35	22	14	40	23	15
	LR	90	18	67	89	20	61
0.05	MH	81	68	59	84	71	61
	LR	98	61	87	97	65	86

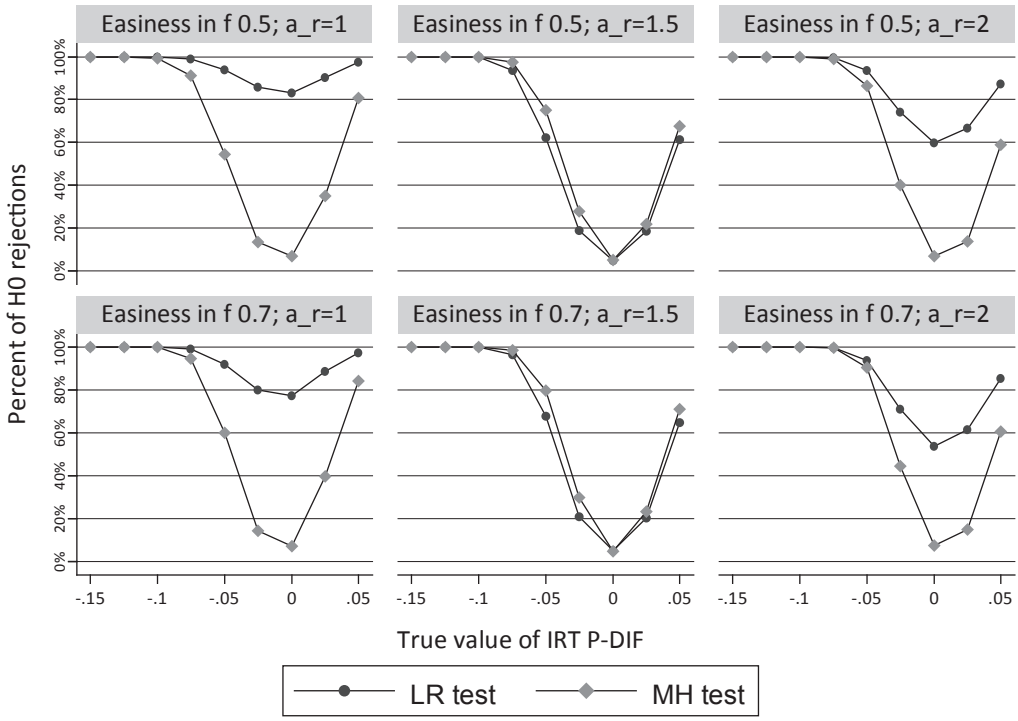

Figure 5. Percentage of replications in which statistically significant result was observed from Mantel–Haenszel and from likelihood ratio test.

resulted in statistically significant values of their respective test statistics over all conditions under study. These values are also visualised in Figure 5.

In cases of a large DIF effect size defined on the item easiness metric (*IRT P – DIF*), ranging from -0.15 to -0.10 both methods concordantly reported significant DIF with probability approaching 1.

The MH test detects more DIF cases that LR under all conditions with $a^r_{20} = 1.5$ (uniform DIF) and the relation reverses for conditions $a^r_{20} = 1$ and $a^r_{20} = 2$. Thus, results of the study illustrate the previously mentioned property of the MH test to be uniformly the most powerful test in the case of validity of the constant odds ratio assumption and confirm the hypothesis that the LR test would be more sensitive in detecting cases of non-uniform DIF.

When DIF is not uniform, sensitivity of the MH test has an interactive dependence on the discrimination parameter and the direction of DIF. For negative values of true *IRT P – DIF* effect, the MH test was more powerful in the case of the higher value of item discrimination ($a^r_{20} = 2$) and for positive values of *IRT P – DIF* effect, the MH test was more powerful in the case of lower discrimination ($a^r_{20} = 1$). Whereas, in almost all of analysed cases of true *IRT P – DIF* effect, the LR test was more powerful in detecting DIF under the condition of the item being less discriminative in the reference group ($a^r_{20} = 1$). The LR test was only slightly more powerful in detecting DIF in case of $a^r_{20} = 2$ than in case of $a^r_{20} = 1$ when the *IRT P – DIF* effect was either 0.05 or -0.075.

Easiness of the 20th item in the focal group is another interesting factor that differentiates

the sensitivity of the methods analysed. When DIF is uniform both tests are more sensitive in detecting DIF than when the item is easier in group *f*, i.e. when its easiness equals 0.7. However when DIF is not uniform this pattern disappears. It is probable, that the relationship is moderated by *IRT P – DIF* effect size, yet closer examination of such interactions would require increasing the number of levels of *IRT P – DIF* effect in the experiment.

The second problem to analyse was the quality of three different estimators of DIF effect size expressed on the scale of item easiness: $MH\widehat{P-}DIF$ (9), $STD\widehat{P-}DIF$(10) and $IRT\widehat{P-}DIF$ (11). Properties of these estimators were verified against a reference value of the true *IRT P – DIF*, which was known by virtue of the values of IRT model parameters used for generating the simulation data. By averaging the results obtained from 10 000 replications the bias and the standard deviation of the three estimators was assessed.

Figure 6 shows the bias of the three estimators for DIF effect conditional on the values of variables that were manipulated in the experiment. The estimator that is based on the common odds ratio α_{MH} exhibits the highest bias which increases in size with an increase of discrimination of the 20th item in the reference group and with increase of the true *P – DIF* effect size. Direction of the bias of $MH\widehat{P-}DIF$ results in overestimation of the absolute value of *P – DIF*. The estimator $STD\widehat{P-}DIF$ is far less biased than $MH\widehat{P-}DIF$ and its bias seems not to depend on the value of a^r_{20} significantly, however, a clear relation between bias and the true value of *P – DIF* is observed. Direction of bias of $STD\widehat{P-}DIF$ is such that it leads to underestimation of the absolute value of

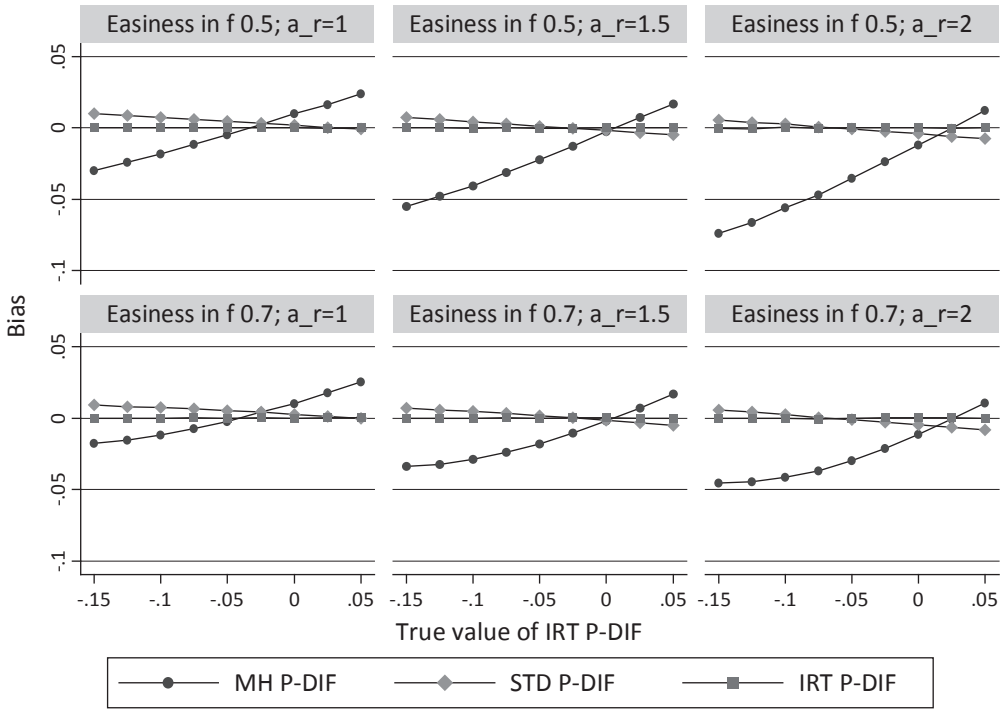

Figure 6. Bias of $P-DIF$ effect size estimators in relation to experimental conditions (horizontal axis represents true value of $P-DIF$, vertical axis stands for bias).

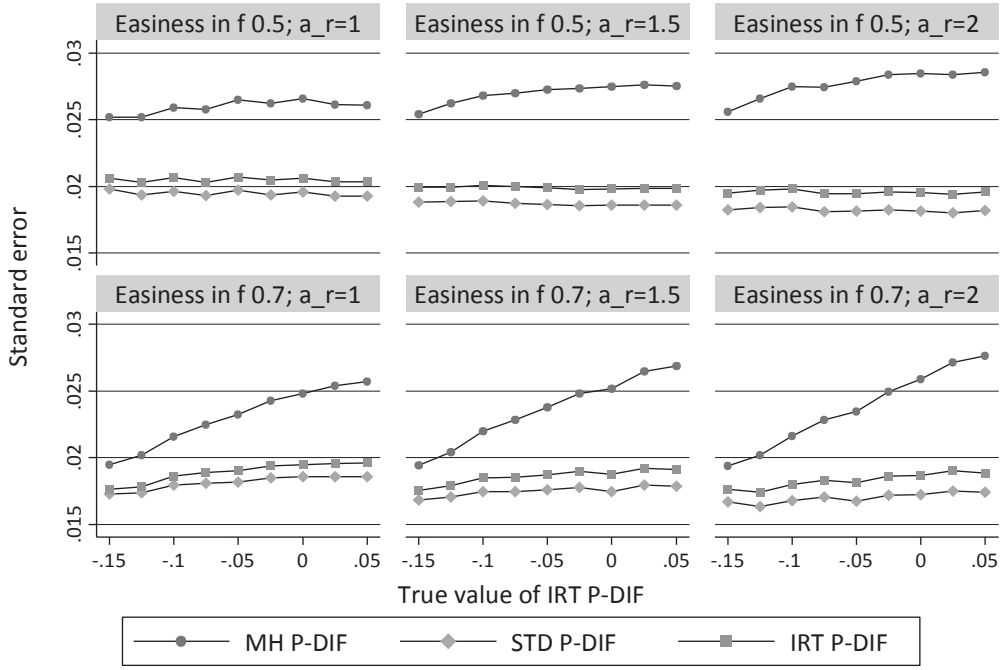

Figure 7. Standard errors of $P-DIF$ effect size estimators in relation to experimental conditions (horizontal axis represents true value of $P-DIF$, vertical axis stands for standard error).

$P-DIF$ effect size – $STD\widehat{P-DIF}$ shrinks the estimates towards zero which is opposite to the behaviour $MH\widehat{P-DIF}$. The $IRT\ P-DIF$ estimator did not reveal bias of any practical significance[2] in any of the experimental conditions studied.

Figure 7 shows analogous information to Figure 6 but plots the standard deviation of $P-DIF$ effect size estimators. Information on standard deviations of $P-DIF$ estimators is shown in Figure 7 and collected in Table 4. These standard deviations can be seen as Monte Carlo estimations of the standard errors of $P-DIF$ estimators. The $MH\widehat{P-DIF}$ estimator is characterised by the largest standard error. Standard errors of $STD\widehat{P-DIF}$ and $IRT\widehat{P-DIF}$ are comparable, however, the first estimator under all conditions tested had a lower standard error that the second. Systematically lower standard error of $STD\widehat{P-DIF}$ in comparison to $IRT\widehat{P-DIF}$ is probably a consequence of differences in bias of these two estimators (Figure 6) – the more an estimator shrinks towards zero the lower its variance.

It can be noticed that standard errors of $P-DIF$ estimators increase when easiness of the 20th item in the focal group is smaller (0.5). Also a negative relation between the 20th item discrimination in the focal group and value of the standard error can be considered for $STD\widehat{P-DIF}$ and $IRT\widehat{P-DIF}$. Their standard errors increase when the discrimination parameter decreases. The relation

[2] It is highly plausible that due to the fact that ML estimators of parameters of an IRT model are biased (Lord, 1983) also an estimator such as $IRT\widehat{P-DIF}$ will be biased because it is constructed by substitution of estimates of IRT model parameters into eq. (11). Kondratek (2012) presented examples of how IRT based estimates of observed score distribution "inherit" bias from estimators of IRT model parameters. The bias of $IRT\widehat{P-DIF}$ under conditions tested in the experiment was yet so small that can be viewed as negligible from practical point of view.

Table 4
Standard errors of $P - DIF$ effect size estimators in relation to experimental conditions (values in the table need to be multiplied by 0.01)

IRT $P - DIF$	Statistic	Easiness of the 20th item in f: 0.5			Easiness of the 20th item in f: 0.7		
		$a^r_{20} = 1$	$a^r_{20} = 1.5$	$a^r_{20} = 2$	$a^r_{20} = 1$	$a^r_{20} = 1$	$a^r_{20} = 1.5$
-0.15	MH	2.52	2.54	2.56	1.95	1.94	1.94
	STD	1.98	1.88	1.82	1.72	1.68	1.67
	IRT	2.06	1.99	1.95	1.76	1.75	1.76
-0.125	MH	2.52	2.62	2.66	2.02	2.04	2.02
	STD	1.94	1.88	1.84	1.74	1.70	1.63
	IRT	2.03	1.99	1.97	1.78	1.79	1.74
-0.1	MH	2.59	2.68	2.75	2.16	2.20	2.16
	STD	1.96	1.89	1.85	1.79	1.74	1.68
	IRT	2.06	2.01	1.98	1.86	1.85	1.80
-0.075	MH	2.58	2.70	2.74	2.25	2.29	2.28
	STD	1.93	1.87	1.81	1.81	1.75	1.71
	IRT	2.03	2.00	1.94	1.89	1.85	1.83
-0.05	MH	2.65	2.72	2.79	2.32	2.38	2.35
	STD	1.97	1.86	1.81	1.82	1.76	1.67
	IRT	2.07	1.99	1.95	1.90	1.87	1.81
-0.025	MH	2.62	2.73	2.84	2.43	2.48	2.49
	STD	1.94	1.85	1.82	1.85	1.78	1.72
	IRT	2.05	1.98	1.96	1.94	1.90	1.86
0	MH	2.66	2.75	2.85	2.48	2.52	2.59
	STD	1.96	1.86	1.81	1.86	1.75	1.72
	IRT	2.06	1.98	1.95	1.95	1.87	1.86
0.025	MH	2.61	2.76	2.83	2.54	2.65	2.71
	STD	1.93	1.86	1.80	1.86	1.80	1.75
	IRT	2.04	1.99	1.94	1.96	1.92	1.90
0.05	MH	2.61	2.75	2.85	2.57	2.69	2.76
	STD	1.93	1.86	1.82	1.86	1.79	1.74
	IRT	2.03	1.98	1.96	1.96	1.91	1.89

between discrimination of the item and magnitude of standard error of $MH\ \widehat{P-DIF}$ is not clear.

The final goal of the experiment was to compare the two DIF effect size classification systems described in the article. Presentation of the results starts from examining the behaviour of ETS's classification based on *MH D – DIF* owing to its popularity. Later the two methods are directly compared.

Table 5
Percentage of replications in which item was classified A, B or C based on MH D – DIF in relation to easiness of item in the focal group and discrimination of item in the reference group

Easiness of the 20th item in *f*	Value of a^r_{20}	*MH D – DIF* classification result		
		A	B	C
0.5	$a^r_{20} = 1$	66	19	15
	$a^r_{20} = 1.5$	59	18	23
	$a^r_{20} = 2$	55	18	27
0.7	$a^r_{20} = 1$	58	17	25
	$a^r_{20} = 1.5$	53	17	31
	$a^r_{20} = 2$	49	16	35

Results of DIF classification based on estimated *MH D – DIF* value (Figure 2) in relation to easiness of the 20th item in the focal group and in relation to discrimination of this item in the reference group are shown in Table 5. Special attention is required when inspecting the category labelled "C" since it is designed to signal items with the largest DIF between the focal and reference groups. There is a noticeable pattern in the number of cases labelled "C", increasing with item discrimination in the reference group or when its easiness increases in the focal group. Thus the classification results are strictly dependent on the discrimination of the item in the reference group and easiness of the item in the focal group, which is probably a reflection of the pattern previously mentioned of the relationship between the statistical power of the MH test and the experimental conditions tested (Table 3 and Figure 5).

Table 6 summarises results of the *MH D – DIF* classification in more detail. The same information is plotted in Figure 8. The previous observation that category C is applicable more often when easiness of an item in the focal group is higher (0.7) is also valid when item discrimination is controlled for. If an item's difficulty in the focal group is held constant, a general tendency is observed that when it is more discriminative for the reference group, it is more probable that it will be allocated a higher (with larger DIF) category.

In order to illustrate that ETS classification more often results in a higher DIF category with increase of item discrimination, cases when easiness in the focal group was

Table 6
Percentage of replications in which item was classified A, B or C based on MH D – DIF in relation to experimental conditions

IRT P – DIF	Easiness of the 20th item in *f*: 0.5									Easiness of the 20th item in *f*: 0.7								
	$a^r_{20} = 1$			$a^r_{20} = 1.5$			$a^r_{20} = 2$			$a^r_{20} = 1$			$a^r_{20} = 1.5$			$a^r_{20} = 2$		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
-0.15	0	14	85	0	3	98	0	0	100	0	0	100	0	0	100	0	0	100
-0.125	4	54	42	0	25	75	0	10	90	0	13	87	0	3	97	0	0	100
-0.1	30	63	7	9	63	28	3	49	48	6	60	33	1	35	64	0	17	83
-0.075	77	23	0	47	50	3	26	65	9	48	49	3	20	67	13	7	63	29
-0.05	97	3	0	89	11	0	76	24	0	90	10	0	72	27	1	53	45	2
-0.025	100	0	0	99	1	0	97	3	0	100	0	0	98	2	0	93	7	0
0	100	0	0	100	0	0	100	0	0	100	0	0	100	0	0	100	0	0
0.025	99	1	0	100	0	0	100	0	0	98	2	0	99	1	0	99	1	0
0.05	88	12	0	92	8	0	93	7	0	78	22	0	85	15	0	89	11	0

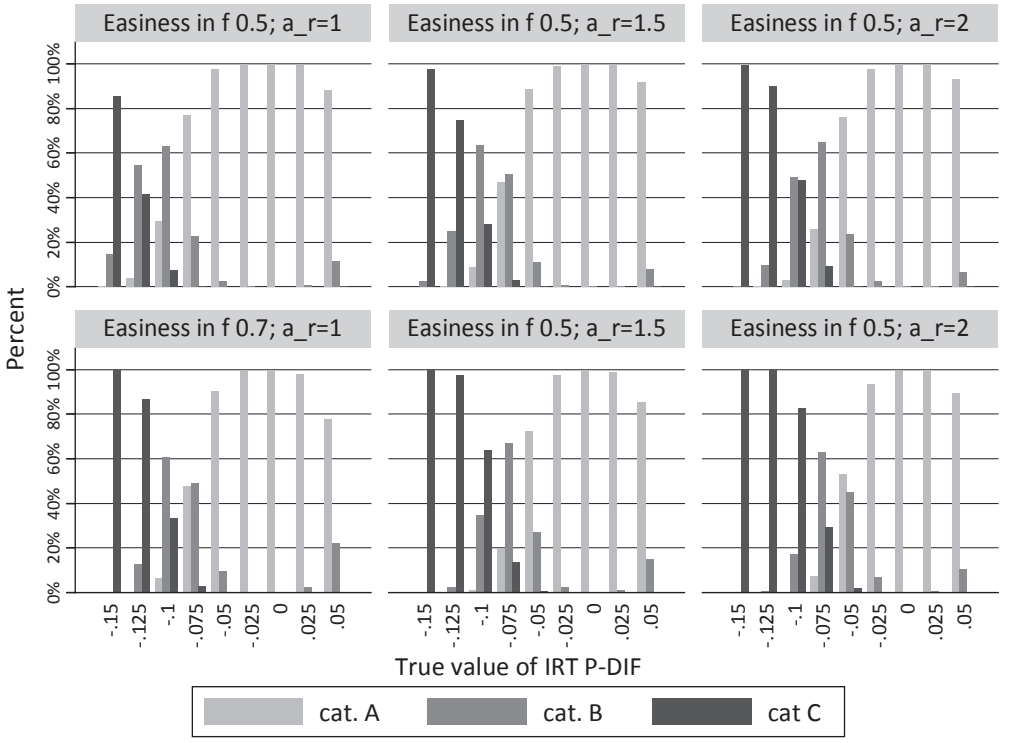

Figure 8. Percentage of replications in which item was classified A, B or C based on *MH D – DIF* in relation to experimental conditions.

0.7 should be considered. When the true *P – DIF* value is set to -0.05 most of the replications result in category A (72% marginal over a^r_{20}) which is a category without DIF. However when inspecting relevant proportions conditional over a^r_{20} we observe large variation with respect to cases in category B – for a discrimination of 1, only 10% of replications fell into this category and for discrimination of 2, as many as 45% replications resulted in B. When a true *P – DIF* value is -0.1, a similar relationship occurs for category C – for = 1, this category applies to 33% of replications, while for $a^r_{20} = 2$ as many as 83% of cases are classified as C.

Graphs in Figure 8 clearly illustrate the previously mentioned (Figure 3) lack of symmetry in how the classification based on *MH D – DIF* treats cases of DIF of similar magnitude but opposite in direction. If the true *P – DIF* effect equals -0.05, the B category is used more often than when its value is 0.05. The ETS classification based on *MH D – DIF*, is therefore more sensitive for cases when an item is easier for the reference group. It might be expected that a scheme for DIF classification should be as effective in detecting DIF in cases when an item favours the reference group as in case when the focal group is favoured.

In conclusion, the experiment showed that results of *MH D – DIF* based DIF effect classification depend on parameters of the item analysed for DIF and the true value of DIF effect size. The scheme was more sensitive if easiness of item in the focal group was larger and if the discrimination of an item in the reference group was greater. Worth

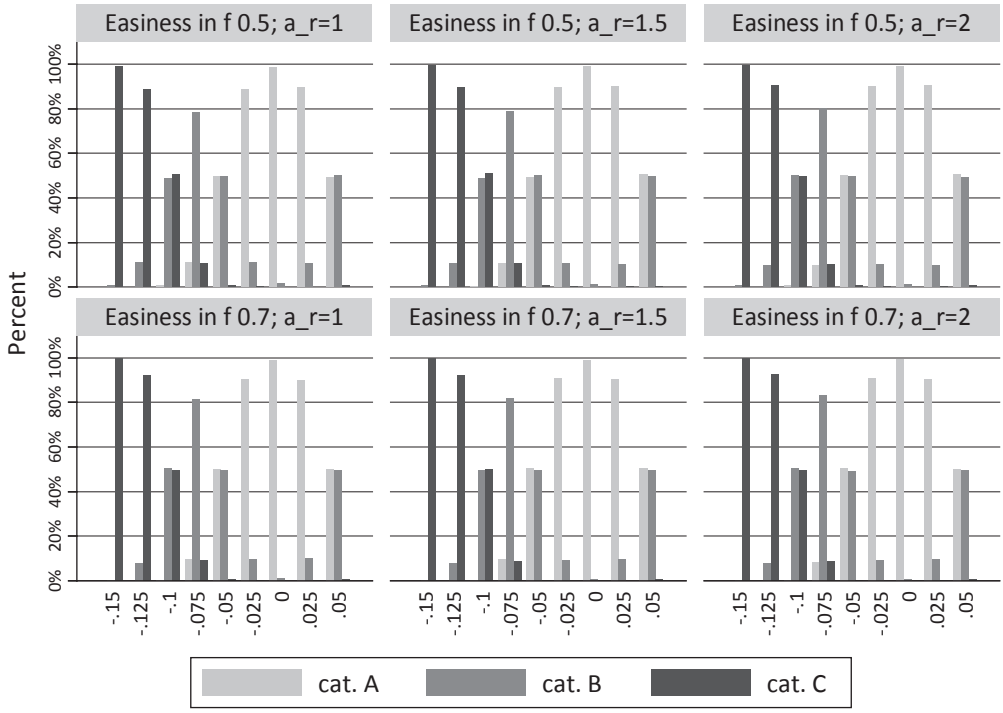

Figure 9. Percentage of replications in which item was classified A, B or C based on *IRT P – DIF* in relation to experimental conditions.

Table 7
Percentage of replications in which item was classified A, B or C based on IRT P – DIF in relation to experimental conditions

IRT P – DIF	Easiness of the 20th item in *f* : 0.5									Easiness of the 20th item in *f* : 0.7								
	$a^r_{20} = 1$			$a^r_{20} = 1.5$			$a^r_{20} = 2$			$a^r_{20} = 1$			$a^r_{20} = 1.5$			$a^r_{20} = 2$		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
-0.15	0	1	99	0	1	99	0	1	99	0	0	100	0	0	100	0	0	100
-0.125	0	11	89	0	11	89	0	10	90	0	8	92	0	8	92	0	8	92
-0.1	1	49	51	0	49	51	1	50	50	0	50	49	0	50	50	0	50	49
-0.075	11	78	11	11	79	10	9	80	10	9	82	9	9	82	9	8	83	9
-0.05	50	50	1	49	50	1	50	49	1	50	50	0	50	50	0	51	49	0
-0.025	89	11	0	89	11	0	90	10	0	90	10	0	91	9	0	91	9	0
0	98	2	0	99	1	0	99	1	0	99	1	0	99	1	0	99	1	0
0.025	89	11	0	90	10	0	90	10	0	90	10	0	90	10	0	90	10	0
0.05	49	50	1	50	49	0	50	49	1	50	50	1	50	49	0	50	50	0

Table 8
Consistency between DIF effect classification schemes based on MH D – DIF and IRT P – DIF [%]

Experimental conditions		Consistent classifications	Inconsistent classifications	Inconsistent classifications					
				P – DIF class = A		D – DIF class = A			
				D – DIF class = C	D–DIF class = B	P – DIF class = C	P – DIF class = B	D – DIF class = B, P – DIF class = C	D – DIF class = C, P – DIF class = B
Easiness of the 20th item in *f* : 0.5	a^r_{20} =1	64	36	0	0	0	23	13	0
	a^r_{20} =1,5	78	22	0	0	0	16	5	0
	a^r_{20} =2	87	13	0	0	0	12	1	1
Easiness of the 20th item in *f* : 0.7	a^r_{20} =1	82	18	0	0	0	15	3	0
	a^r_{20} =1,5	88	12	0	0	0	9	0	3
	a^r_{20} =2	86	14	0	1	0	6	0	7

emphasizing is the lack of symmetry of the procedure: ETS classification is more efficient in detecting DIF in cases of items that favour the reference group.

Figure 9 and Table 7 present the distribution of A, B, C categories observed over all experimental conditions for DIF effect classification based on *IRT P – DIF* in an analogous manner as Figure 8 and Table 6 did for *MH D – DIF* based classification, so allowing comparisons. Corresponding distributions of classifications are similar, however when inspected closer it appears that *MH D – DIF* classification fluctuates more following changes of an item's easiness in focal group and an item's discrimination in the reference group, whereas *IRT P – DIF* classification is very stable over all experimental conditions.

In Table 8 the two classification schemes are compared with respect to their consistency. In most cases the two schemes classified items into the same categories and the percentage of consistent classifications was positively related to increased item discrimination in the reference group and to increase of easiness of item in the focal group. Inconsistent classifications were divided into the case when both schemes led to conclusions that DIF was present but differed with regard to its degree and the case when one of the schemes placed an item in category A while the other pointed to category B or C. It is worth noting that inconsistent classifications generally followed the pattern that an item was more likely to be classified lower by the *MH D – DIF* than by *IRT P – DIF* classification.

Conclusions

The article aimed to compare two tools applied to DIF detection: the Mantel–Haenszel test and an approach based on the likelihood ratio test of parametric IRT models. The Monte Carlo experiment that was conducted, allowed for verification of performance of the two methods in various experimental conditions.

Results confirmed that Mantel–Haenszel test is more powerful in detecting uniform DIF, however it reduces in power when an interaction between magnitude of DIF and ability level is present. In cases of non-uniform

DIF fitting a two-parameter logistic IRT model followed by conducting a likelihood ratio test was a more powerful approach, as the IRT model allowed for interaction between DIF magnitude and ability level explicitly.

It was discovered that DIF effect size estimated on the scale of easiness of the item by IRT modelling resulted in bias that was negligible from a practical point of view. Moreover standard errors for the classical $STD\widehat{P-DIF}$ estimator of the same parameter were comparable to standard errors of the $IRT\widehat{P-DIF}$ estimator. This makes it possible to use estimates of standard errors of $STD\widehat{P-DIF}$, for which formulas are available (Dorans and Holland, 1993), as an approximation of standard errors of *IRT P – DIF*. Such a solution allows augmentation of the *IRT P – DIF* classification, adding information about precision of estimating $STD\widehat{P-DIF}$ in an analogous manner to the *MH D – DIF* classification.

Literature

Agresti, A. (2002). *Categorical data analysis*. New Jersey: John Wiley & Sons.

Dorans, N. J. and Holland, P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland and H. Wainer (eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Earlbaum.

Glas, C. A. (2010). *Preliminary manual of the software program Multidimensional Item Response Theory (MIRT)*. Enschede: University of Twente.

Kondratek, B. (2012). *Bias of IRT observed score equating under NEAT design*. Poster presented at the conference Modern Modelling Methods, Storrs, Connecticut.

Lord, F. M. (1983). Statistical bias in maximum likelihood estimators of item parameters. *Psychometrika*, *48*(3), 425–435.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison–Wesley.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748

Monahan, P. O., McHorney, C. A., Stump, T. E. and Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*(1), 92–109.

Penfield, R. D. and Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao and S. Sinharay (eds.), *Handbook of statistics, Vol. 26. Psychometrics* (pp. 125–167). New York, NY: Elsevier.

Radhakrishna, S. (1965). Combination of results from several 2×2 contingency tables. *Biometrics*, *21*(1), 86–98.

Swaminathan, H. and Rogers J. H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

Thissen, D., Steinberg, L. and Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland and H. Wainer (eds.), *Differential Item Functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Earlbaum.

Wainer, H. (1993). Model-based standardized measurement of an items differential impact. In P. W. Holland and H. Wainer (eds.), *Differential Item Functioning* (pp. 255–276). Hillsdale, NJ: Lawrence Earlbaum.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics*, *19*(4), 251–253.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland and H. Wainer (eds.), *Differential Item Functioning* (pp. 337–348). Hillsdale, NJ: Lawrence Earlbaum.

Zieky, M. (2003). *A DIF primer*. Princeton, NJ: Educational Testing Service.