Journal of Educational and Behavioral Statistics
2015, Vol. 40, No. 2, pp. 111–135
DOI: 10.3102/1076998614559747
© 2014 AERA. http://jebs.aera.net

Detection of Differential Item Functioning Using the Lasso Approach

David Magis
KU Leuven
University of Liège

Francis Tuerlinckx
University of Leuven

Paul De Boeck
Ohio State University
University of Leuven

This article proposes a novel approach to detect differential item functioning (DIF) among dichotomously scored items. Unlike standard DIF methods that perform an item-by-item analysis, we propose the "LR lasso DIF method": logistic regression (LR) model is formulated for all item responses. The model contains item-specific intercepts, an effect of the sum score, and item-group interaction (i.e., DIF) effects, with a lasso penalty on all DIF parameters. Optimal penalty parameter selection is investigated through several known information criteria (Akaike information criterion, Bayesian information criterion, and cross validation) as well as through a newly developed alternative. A simulation study was conducted to compare the global performance of the suggested LR lasso DIF method to the LR and Mantel–Haenszel methods (in terms of false alarm and hit rates). It is concluded that for small samples, the LR lasso DIF approach globally outperforms the LR method, and also the Mantel–Haenszel method, especially in the presence of item impact, while it yields similar results with larger samples.

Keywords: *differential item functioning; Rasch model; penalized maximization; lasso; information criterion*

Introduction

The identification of differential item functioning (DIF) of dichotomous items is an important field of current research. A variety of methods has been proposed. The main two categories are test score–based methods and item response theory (IRT) modeling methods. Methods of the first type make use of a matching variable, the most popular of these methods being the Mantel–Haenszel (Holland & Thayer, 1988), standardization (Dorans & Kulick, 1986), logistic regression (LR; Swaminathan & Rogers, 1990), and SIBTEST (Shealy & Stout, 1993). Methods of the second type rely on differences in item parameters as estimated with an IRT model, such as Lord's χ^2 test (Lord, 1980), Raju's (1988, 1990) area approach, and the likelihood ratio test (Thissen, Steinberg, & Wainer, 1988). Except for a few of them, methods of both types have in common that a DIF index or statistic is derived for each item separately and that items are flagged as DIF if this statistic exceeds a detection threshold based on statistical or practical considerations.

Computing DIF statistics separately for each item comes with two possible problems: multiple testing and a contaminated anchor set. First, when more than 1 item is tested, the DIF detection threshold should be adjusted for multiple testing in order to avoid overidentification of DIF. Recently, Kim and Oshima (2013) investigated several multiple testing adjustments in the DIF framework, including the Bonferroni correction and the Benjamini–Hochberg false discovery rate. Such a multiple testing adjustment, though having established some interest (Penfield, 2001; Thissen, Steinberg, & Kuang, 2002), has not yet become of systematic use.

The second problem of an item-by-item approach is the assumption that all items except the one under investigation (i.e., the anchor set) are considered to be DIF free, which is not guaranteed. If all items are tested, this assumption is made regarding all subsets of size $J - 1$, where J is the number of items. This leads to the paradoxical situation that, in order to apply an item-by-item approach to all items to detect DIF, one must assume that there is no DIF. This assumption becomes problematic when true DIF items are present in the data, since they may strongly influence the DIF statistics and lead to an inaccurate identification of DIF items (leading to a contaminated anchor set). As a consequence, one often observes an inflation of the Type I error rates (e.g., Clauser, Mazor, & Hambleton, 1993; Magis & De Boeck, 2012; Wang & Su, 2004). The most common approach to overcome this issue is an iterative procedure, often referred to as *item purification*, in order to reduce the impact of other DIF items on the DIF statistics for an item under investigation (Candell & Drasgow, 1988). More recently, two different alternatives for this iterative process have been suggested: an item mixture model (Frederickx, Tuerlinckx, De Boeck, & Magis, 2010) and an outlier approach based on robust statistics (Magis & De Boeck, 2011, 2012).

In this article, a regularization method is proposed to look at all items simultaneously. It is based on an LR approach with test scores as the matching variable, in a model with item difficulty parameters (item main effects) as well as DIF parameters (item–group interaction terms). More precisely, the lasso penalization is chosen for the regularization, allowing the item difficulty levels to be estimated freely but constraining the interaction terms (i.e., the DIF parameters) under the lasso shrinkage condition (Hastie, Tibshirani, & Friedman, 2009;

Tibshirani, 1996). This approach is further referred to as the LR *lasso DIF* (or *LR lasso DIF* or shortly *lasso DIF* when there is no ambiguity) method and solves the two problems mentioned earlier. First, it is a method for all items simultaneously so that multiple testing is not a problem. Second, no assumptions are made regarding which subset of items should be DIF free. Interestingly, a somewhat similar approach was developed independently and in a somewhat different DIF framework (Tutz & Schauberger, in press). The main difference, however, is that Tutz and Schauberger's approach makes use of a latent variable model with the ability level of the respondent as latent trait, as in usual IRT modeling. Our approach relies on a simpler LR by making use of a proxy for the ability level, that is, the test score (in line with previous DIF approaches as e.g., Swaminathan & Rogers, 1990).

This article has the following three main objectives: (a) to describe the LR lasso DIF method, (b) to develop tools for optimal lasso penalty parameter selection, and (c) to compare the overall performance of the LR lasso DIF method to other DIF methods. We will conduct a simulation study, and the LR (Swaminathan & Rogers, 1990) method will be used as the reference standard DIF method, as it is the natural standard DIF method from which the present method is derived. For benchmarking reasons, we have also compared the LR lasso DIF with the Mantel–Haenszel (M-H) method (Holland & Thayer, 1988).

DIF and Lasso

This section describes first the DIF framework and the related underlying modeling. Then the penalized lasso regression approach is described and applied to this DIF context.

Framework

We consider a test of J dichotomously scored items that are presented to two groups of respondents, the *reference group* and the *focal group*, which are further denoted by subscripts 0 and 1, respectively. The sizes of these groups are N_0 and N_1, respectively, and N is the total sample size. Let Y_{ijg} be the response of examinee i ($i = 1, \ldots, N_g$) from group g ($g = 0, 1$) to item j ($j = 1, \ldots, J$), coded as 0 for an incorrect response and as 1 for a correct response.

This article focuses on the identification of uniform DIF by means of test score–based methods. More precisely, the starting point is the LR method (Swaminathan & Rogers, 1990), which aims at modeling the (logit of the) probability of answering the item correctly as a function of the group membership and the test score:

$$\text{Logit}\left[\Pr(Y_{ijg} = 1)\right] = \alpha_{0j} + \alpha_{1j}S_i + \alpha_{2j}G_{ig}, \ j = 1, \ldots, J, \tag{1}$$

where S_i is the test score of respondent i and G_{ig} is the group membership indicator, being equal to 1 if respondent i belongs to the reference group

($g = 1$) and 0 otherwise. Parameters $\alpha_{01}, \ldots, \alpha_{0J}$ are further referred to as the item difficulty parameters since they are the counterparts of the usual item difficulties in an IRT framework. Similarly, parameters $\alpha_{11}, \ldots, \alpha_{1J}$ can be considered as the counterparts of item discrimination parameters. Finally, parameters $\alpha_{21}, \ldots, \alpha_{2J}$ are the parameters of interest for the DIF approach as will be explained.

Model 1 states that the probability of answering the item correctly depends on the item, on the test score (which is used in this framework as a proxy for the ability level) and the group membership. This model is being fitted separately for each item, and the G_{ig} terms stand for the uniform DIF effect. Testing for uniform DIF therefore requires testing for statistical significance of each coefficient α_{2j}, which is usually performed by a Wald test or a likelihood ratio test by comparing Model 1 to the simpler model with the test score as single covariate. The parameters $\alpha_{21}, \ldots, \alpha_{2J}$ are further referred to as the *DIF parameters*.

As stated previously, there are two drawbacks related to this method. First, by fitting Model 1 separately for each item, one enters into the previously mentioned multiple testing issue that was pointed out by Kim and Oshima (2013) among others. Second, all other items than the item under investigation are considered as DIF-free items. These drawbacks can be solved in a more general approach that consists in fitting a single general model that contains all DIF parameters (and thus, removing the multiple testing issue).

For our purpose, the effect of the test score (modeled by coefficients α_{1j}) is constrained to be identical for all items for two reasons. First, the sum score as a proxy for the ability, as in the LR approach, is in fact more in line with the one-parameter logistic model (e.g., del Pino, San Martin, Gonzalez, & De Boeck, 2008) than with the two-parameter logistic model. Strictly speaking the latter model would require a weighted sum score as a covariate, while the weights for that sum score are of course unknown. Second, allowing for differences in item discrimination can lead to a rather high rate of false DIF detection if the groups differ with respect to their ability level (DeMars, 2010; Magis & De Boeck, in press). In sum, the present constraint on α_{1j} coefficients avoids a theoretical inconsistency and potentially misleading results.

Therefore, our modeling approach yields:

$$\text{Logit}\left[\Pr(Y_{ijg} = 1)\right] = \alpha_{0j} + \alpha_1 S_i + \alpha_{2j} G_{ig}, \tag{2}$$

and all item parameters are estimated simultaneously in a single modeling approach. Once fitted, uniform DIF can be investigated by examining the DIF parameters α_{2j} as usual. Note that the main effect of group membership is not introduced in Equation 2 to avoid that one of the α_{2j} needs to be constrained to 0. Furthermore, because the test score is a sufficient statistic for ability in this Rasch modeling context, we anticipate that the test score as a covariate can already capture the group effect (as a main effect).

Penalized Lasso LR

Let us collect the model parameters in a single vector $\boldsymbol{\tau} = (\alpha_{01}, \ldots, \alpha_{0J}, \alpha_1, \alpha_{21}, \ldots, \alpha_{2J})$. Maximum likelihood estimation proceeds by determining $\hat{\boldsymbol{\tau}}_{\mathbf{MLE}}$, the set of parameter values that maximizes the log likelihood $l(\boldsymbol{\tau})$. In a Rasch (1960) modeling context, joint estimation of all model parameters is expected to have minor effects when the ability levels are replaced by the test score as a proxy (del Pino et al., 2008). Note however that the maximum likelihood estimator is unidentified in this problem, unless an additional constraint is added to the DIF parameters (e.g., $\alpha_{21} = 0$). Penalized lasso maximization (Hastie et al., 2009) will be applied to determine on the DIF items.

In the penalized LR lasso DIF approach, we add a penalization term for the DIF parameters $(\alpha_{21}, \ldots, \alpha_{2J})$ to the log likelihood (while leaving both the score parameter α_1 and the item difficulty parameters $\alpha_{01}, \ldots, \alpha_{0J}$ unaffected). The parameter estimates are found by maximizing the penalized log likelihood:

$$\hat{\boldsymbol{\tau}}(\boldsymbol{\lambda}) = \arg\max\ l(\boldsymbol{\tau}) - \lambda \sum_{j=1}^{J} |\alpha_{2j}|, \tag{3}$$

where the second term on the right-hand side is the penalty term.

The λ parameter in Equation 3 is the *penalty parameter* and it determines the balance between the goodness-of-fit term (i.e., $l(\boldsymbol{\tau})$) and the penalty term. Smaller values of λ downweight the penalty and lead to parameter estimates close to the maximum likelihood estimates (with perfect equality whenever $\lambda \to 0$), while larger values of λ enlarge the penalty and lead to a general decrease in the DIF parameters α_{2j} toward 0. The lasso penalty has indeed the attractive feature that it drives the DIF parameters effectively to 0 for an increasing value of λ (see Hastie et al., 2009), in contrast to other penalization approaches, such as ridge regression, which only makes all penalized coefficients smaller in size with a larger penalty.

Selection of an Optimal Penalty Parameter

Although the application of the lasso penalty to this DIF context may seem clear from Equation 3, an important issue remains unresolved: Which value of λ is optimal? A particular value of λ coincides with a specific configuration of values for the DIF parameters α_{2j}, which are consequently further denoted by $\alpha_{2j}(\lambda)$. Following the lasso approach, the non-DIF items should approach 0 earlier than the DIF items when λ goes to 0. In other words, the issue is what the value of λ should be for an optimal differentiation between DIF items and non-DIF items.

These ideas are illustrated in Figure 1, which is based on a simulated data. A test of 20 items was created with item difficulties drawn from a standard normal distribution. Two groups of 500 respondents were considered, and proficiency

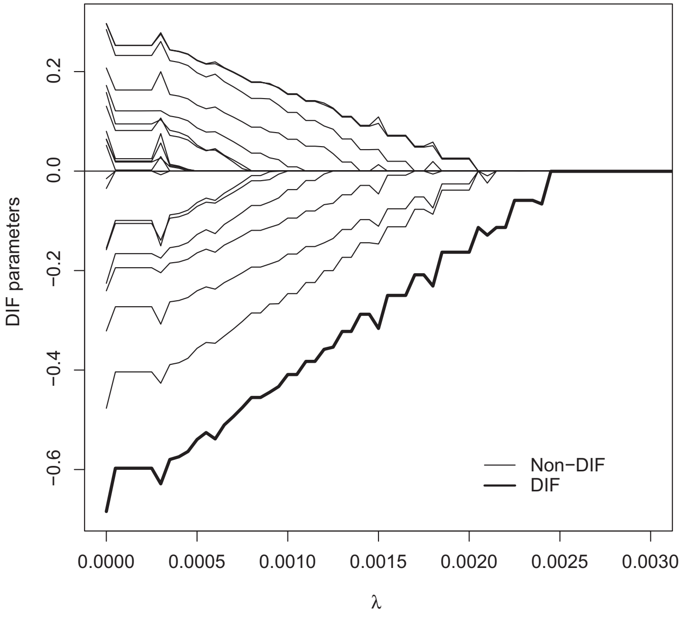

FIGURE 1. *Regularization paths of differential item functioning parameters for a sequence of* λ *values, artificial data set.*

levels of the respondents were drawn from a standard normal distribution in both groups (without item impact). Item responses were generated using the one-parameter logistic model (see next section for further details). One item was selected to function differently by increasing its difficulty level in the focal group with a value of 0.5. Model parameters were then estimated under lasso penalty by maximizing Equation 3 for a sequence of penalty parameters λ (using the *glmnet* function of the eponym R package; see Friedman, Hastie, & Tibshirani, 2010).

Figure 1 displays the estimated DIF parameters $\alpha_{2j}(\lambda)$ as functions of the penalty parameter λ for all items. Tracing out $\alpha_{2j}(\lambda)$ as a function of λ is called the *regularization path* of $\alpha_{2j}(\lambda)$. The DIF item has a thicker curve than the other, non-DIF items. At $\lambda = 0$ (hence, no lasso penalty), the DIF item does already depart from the other non-DIF items by a much larger value of its DIF parameter estimate. As λ increases, some DIF parameters quickly decrease toward 0, while other parameters require larger values of the penalty parameter to reach 0, thus indicating a stronger DIF effect. As expected, a larger λ leads to less nonzero DIF parameters.

The selection of the optimal value λ* of λ is a central and crucial aspect of the approach. Selecting a too small λ* value will increase the Type I error rate (taking non-DIF for being DIF), while a too large λ* value is conservative and will lead to an increase of the Type II error rate (too few items being flagged as DIF). Because a particular λ-value balances fit (as indicated by the log likelihood) and complexity (indicated through the number of nonzero DIF parameters), the selection of an optimal λ parameter can be framed as a model selection problem that has desirable generalization properties (Pitt & Myung, 2002). As a result, we can use model selection tools for the selection of a λ^* value. In this article, we consider two such approaches: cross validation (CV) and information criteria.

CV (Hastie et al., 2009, pp. 241–245) consists in splitting the data set in K subsets or folds, fitting the model repeatedly on $K - 1$ folds (by successively removing one of the folds) and then by computing the prediction error of the fitted model when applied to the left-out fold. Prediction errors are then accumulated over the K iterations and the final λ^* value is selected as the one that minimizes the overall prediction error. In the current framework, the splitting is made across respondents, groups and items, with the idea that one fold holds a sample of respondents from both groups and for each respondent a subset of item responses are selected, in such a way that all model parameters can still be estimated in the CV process. The usual deviance statistic is used to compute the prediction error for the lasso penalized LR Model 2. In the following, three numbers of folds are considered ($K = 3$, 5 or 10) and the corresponding optimal λ^* selection methods are referred to as CV3, CV5, and CV10, respectively.

The second strategy consists in considering some well-known information criteria, such as the Akaike information criterion (AIC; Akaike, 1973):

$$\lambda^*_{\text{AIC}} = \arg\min \text{AIC}(\lambda) = \arg\min(-2l(\hat{\boldsymbol{\tau}}(\lambda)) + 2K(\hat{\boldsymbol{\tau}}(\lambda))), \tag{4}$$

and the Bayesian information criterion (BIC; Schwarz, 1978):

$$\lambda^*_{\text{BIC}} = \arg\min \text{BIC}(\lambda) = \arg\min(-2l(\hat{\boldsymbol{\tau}}(\lambda)) + K(\hat{\boldsymbol{\tau}}(\lambda)) \log n), \tag{5}$$

where $K(\hat{\boldsymbol{\tau}}(\lambda))$ is the number of parameters in the model and n is the total number of item responses in the data. Although there is some debate on how $K(\hat{\boldsymbol{\tau}}(\lambda))$ should be calculated, a common choice is to use for $K(\hat{\boldsymbol{\tau}}(\lambda))$ the cardinality of the parameter vector $\boldsymbol{\tau}$ (i.e., the number of nonzero parameters): $K(\hat{\boldsymbol{\tau}}(\lambda)) = \text{card}(\hat{\boldsymbol{\tau}}(\lambda))$ (Tutz & Schauberger, in press; Yuan & Lin, 2006).

Let us illustrate this AIC and BIC approach with the previously described artificial example. For a regular sequence of λ values ranging from 0 to 0.003, the AIC (λ) and BIC (λ) were computed and are displayed in Figure 2. For the AIC criterion (left panel), the optimal penalty parameter λ^* equals 0.0005, while for the BIC criterion (right panel), it equals 0.0024. For this example, the AIC has a higher hit rate (more true DIF items flagged as DIF) than the BIC but also a higher Type I error rate (more non-DIF items flagged as DIF).

Both the results from CV and optimal information criteria (using AIC and BIC) are summarized in Figure 3. This figure displays exactly the same DIF parameters as in Figure 1, but in addition the optimal λ^* values are represented

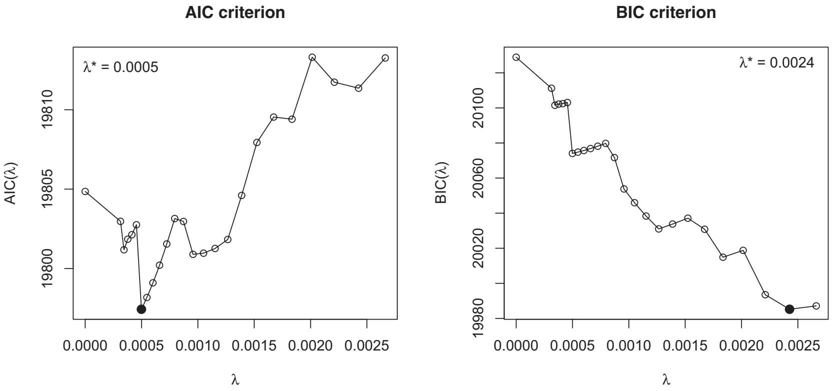

FIGURE 2. *Values of AIC(λ) and BIC(λ) criteria for a sequence of λ values for the artificial data set as in Figure 1. The optimal λ* parameter and related values AIC(λ*) and BIC(λ*) of the criteria are represented by a full black dot. AIC = Akaike information criterion; BIC = Bayesian information criterion.*

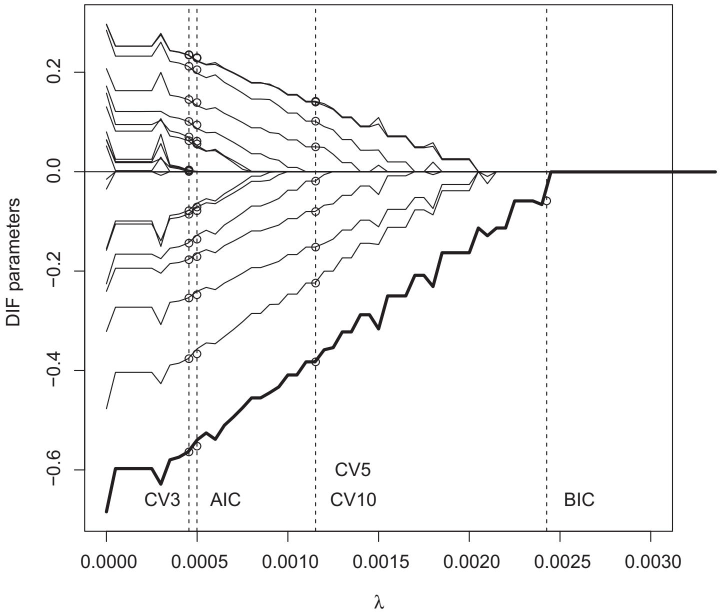

FIGURE 3. *Regularization paths of differential item functioning (DIF) parameter estimates for a sequence of λ values and identification of DIF items with AIC, BIC, and cross-validation, artificial data set. AIC = Akaike information criterion; BIC = Bayesian information criterion.*

with vertical lines. Items flagged as DIF can easily be identified, as their DIF parameter trace lines intersect with the vertical lines. The intersection point is indicated with circles. All methods identify the true DIF item correctly, which was expected from Figure 1, but the AIC and CV3 also led to many false alarms because of a too small λ* value (13 out of 19 and 16 out of 19 false alarms for the AIC and CV3, respectively). The BIC, on the other hand, correctly identifies the one DIF item and does not lead to any false alarm. The CV5 and CV10, for which the λ* value is identical, act in an intermediate way and detect 8 out of the 19 non-DIF items as DIF.

This simple illustration shows that the BIC is more conservative than the AIC for an optimal choice of λ and this finding is in line with previous results (Weakliem, 1999). Some other criteria have been suggested in the literature, such as the AIC3 criterion (Bozdogan, 1993) and the constrained AIC criterion (CAIC; Bozdogan, 1987). Those criteria differ from the AIC and BIC criteria only by their penalties to the log likelihood term and were developed in specific contexts wherein both AIC and BIC failed to perform adequately.

In the current framework, however, some intermediate criterion between the two extreme ones, the conservative BIC and the liberal AIC criteria, is advised. The suggested criterion is further referred to as the *weighted information criterion* (WIC). It is a weighted average approach between the AIC and the BIC criteria. It can be summarized in the following steps:

a. Determine a sufficiently dense grid of weights $\{\omega_i, i = 1, \ldots, K\}$ on the interval $[0, 1]$ (such that 0 coincides with BIC and 1 with AIC).
b. For each weight ω_i, consider the so-called *weighted information criterion*:

$$\mathrm{WIC}(\lambda; \omega_i) = \omega_i \mathrm{AIC}(\lambda) + (1 - \omega_i)\mathrm{BIC}(\lambda) = -2l(\hat{\boldsymbol{\tau}}(\lambda)) + K(2\omega_i + (1 - \omega_i)\log n). \quad (6)$$

See Wu, Chen, and Yan (2013) and Wu and Sepulveda (1998) for a related approach (but in a different context).

c. For each weight ω_i, compute the optimal penalty parameter $\lambda_{\mathrm{WIC}}(\omega)$, which depends on the weight value, by minimizing $\mathrm{WIC}(\lambda; \omega_i)$ with respect to λ: $\lambda_{\mathrm{WIC}}(\omega_i) = \arg\min \mathrm{WIC}(\lambda; \omega_i)$.
d. Select as final optimal penalty parameter λ^*_{WIC}, the median of all unique $\lambda_{\mathrm{WIC}}(\omega_i)$ values over all selected weights.

By allowing the AIC and BIC criteria to be weighted differently depending on the data set, more flexibility is introduced. It is expected that only a finite number of optimal penalty parameters $\lambda_{\mathrm{WIC}}(\omega)$ will be obtained even when a very large number of weight values from 0 to 1 is tried out. If the process returns a unique value for $\lambda_{\mathrm{WIC}}(\omega)$, this means that the AIC and the BIC criteria, obtained by setting, respectively, ω to 1 and 0, return the same optimal penalty parameter and hence, the same classification of items as DIF and non-DIF. When the process leads to only two different values of $\lambda_{\mathrm{WIC}}(\omega)$, then

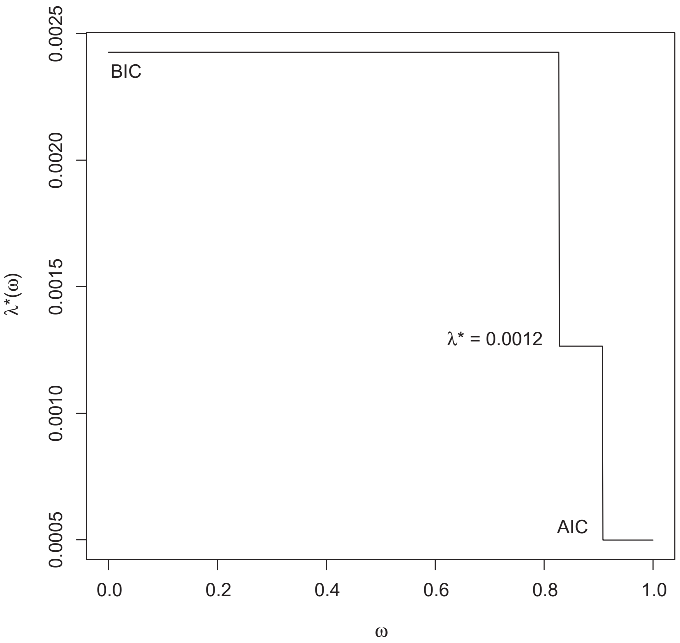

FIGURE 4. *Set of optimal λ^* values for a sequence of weights ω of the WIC(ω) criterion. WIC = weighted information criterion.*

one corresponds to the AIC criterion and the other to the BIC criterion, and the λ^* parameter is the simple average of both values. Otherwise, for more than two different values of $\lambda_{\text{WIC}}(\omega)$, λ^*_{WIC} will lie in between the optimal values obtained with the AIC and the BIC criteria, just as requested.

This weighted average approach is graphically illustrated in Figure 4 for the simulated data example. A sequence of weight values from 0 to 1 by steps of 0.001 was considered, and the optimal penalty values $\lambda_{\text{WIC}}(\omega)$ were computed and plotted against ω. As expected, when ω is close to 0, the optimal penalty parameter of the BIC criterion is obtained (i.e., 0.0024) and for weight values close to one, the optimal weight of the AIC criterion (i.e., 0.0005) is obtained. In between, for weights ranging roughly between 0.8 and 0.9, a different optimal penalty parameter value is selected, $\lambda^*_{\text{WIC}} = 0.0012$. This value is the median value of the three different optimal weights (0.0005, 0.0012, and 0.0024) in this example, and it corresponds therefore to the optimal value based on the WIC criterion Equation 6.

The WIC criterion has a sound empirical basis and is a compromise between the AIC and BIC criteria. Note that the optimal penalty will depend on the characteristics of the problem at hand (such as the test length and the sample size). In order to have a better view on the various methods, a simulation study was conducted and is described in the next section.

Simulation Study

The aim of the study is twofold (a) to compare the different methods for optimal penalty parameter selection and (b) to evaluate the global performance (in terms of false alarm and hit rates) of the LR lasso DIF approach in comparison with the LR and M-H approaches. LR is a natural competitor in this framework, while the M-H method is considered for reasons of benchmarking. Both Objectives (a) and (b) are achieved through a common study design that is described subsequently.

Design

Five factors were manipulated (a) the test length, (b) the group sizes of the reference and focal groups, (c) the item impact, (d) the percentage of DIF items, and (e) the size of DIF. They are described hereafter.

Three test lengths were considered, with 20 items, or 40 items, or 60 items. Three different group sizes were taken into account for both groups: 100, 500, and 1,000 respondents. All combinations of (reference, focal) group sizes were allowed, with the constraint that the focal group size was never larger than the reference group sizes. This led to six pairs, namely, (100, 100), (500, 100), (1000, 100), (500, 500), (1000, 500), and (1000, 1000). That the focal group is smaller than the reference group is not uncommon, and it is known that this can affect the result of DIF analyses (e.g., Penfield, 2001).

The parent distributions for the proficiency levels θ_{ig} of the respondents were selected as follows: in the reference group (i.e., when $g = 0$), θ_{i0} were sampled from a $N(0, 1)$ distribution, while in the focal group (when $g = 0$), θ_{i1} were sampled from a $N(\gamma, 1)$ distribution. Two values of γ were considered, namely, $\gamma = 0$, for the absence of item impact, and $\gamma = -1$, for item impact of one unit in favor of the reference group.

Item difficulties, $\beta_j (j = 1, \ldots, J)$ were drawn from a standard normal $N(0, 1)$ distribution. Three percentages of DIF items were considered, that is, 0%, 5%, and 10%. The first case corresponds to the absence of DIF and is a baseline situation to check the accuracy of the Type I error rates. In the presence of DIF, two DIF sizes δ were selected, $\delta = 0.4$ and $\delta = 0.8$, corresponding to the difference in item difficulty levels between the two groups, in line with previous simulation studies (e.g., Magis & De Boeck, 2012). Note that although DIF was generated asymmetrically (i.e., all DIF items favor the same group), the LR lasso DIF allows for modeling symmetric DIF as well.

This design yields 180 settings, namely, 36 without DIF (3 test lengths, 6 combinations of group sizes, and 2 cases for item impact), and 144 with DIF (by considering in addition 2% of DIF and two DIF sizes). In each setting, 100 data sets were generated. Item responses were drawn as follows. First, for each respondent *i* in group *g* and each item *j*, the (true) probability of answering the item correctly was computed under a Rasch model:

$$P_j(\theta_{i0}) = \frac{\exp(\theta_{i0} - \beta_j)}{1 + \exp(\theta_{i0} - \beta_j)} \text{ or } P_j(\theta_{i_1}) = \frac{\exp(\theta_{i1} - \beta_j - \delta)}{1 + \exp(\theta_{i1} - \beta_j - \delta)}, \tag{7}$$

depending on whether the respondent belongs to the reference or focal group, respectively. In the absence of DIF, δ is set to 0. Second, item responses Y_{ijg} are drawn from a Bernoulli distribution, with success probabilities given by $P_j\left(\theta_{ij}\right)$ as defined in Equation 7.

DIF Methods and Summary Statistics

Finally, eight DIF detection methods were considered. The first one is the LR procedure and the DIF statistic is the likelihood ratio statistic between the two nested models (see also Magis, Béland, Tuerlinckx, & De Boeck, 2010). The second one is the M-H method. The other six methods are based on the lasso-penalized approach and make use of different techniques for optimal penalty parameter selection. Among these techniques, three come from the CV approach, with, respectively, 3, 5, or 10 folds (CV3, CV5, and CV10), and three come from the optimal information criterion framework, that is, the AIC, the BIC, and the WIC criteria.

The identification of DIF was then performed by applying each of the eight methods to the 100 data sets of each simulation setting. False alarm rates (i.e., Type I error) and hit rates (i.e., power values, one minus Type II error rates) were computed per DIF method and simulation setting. The false alarm rate is the proportion of non-DIF items that are incorrectly flagged as DIF, while the hit rate is the proportion of DIF items that are correctly flagged as DIF. Of course, the hit rate cannot be computed when there is no DIF.

Receiver–Operating Characteristic (ROC) Curves, Areas, and Perpendicular Distances

Two types of comparisons of the six lasso-based methods with the LR approach (and with the M-H approach) were performed: a global comparison using ROC curves (Hastie et al., 2009), only for the DIF cells in the design, and a specific comparison for each of the six different ways of determining a lasso penalty. The specific comparisons make use of perpendicular distances from these ROC curves to the identity line, as will be explained.

The reason for using ROC curves is that the performance of the LR and M-H approaches depend on the selection of a significance level, while for the LR lasso DIF method the lasso penalty parameter is the crucial aspect instead. This complicates a comparison in terms of false alarm and hit rates, a common practice in DIF studies (e.g., Clauser et al., 1993; Finch & French, 2007; Frederickx,

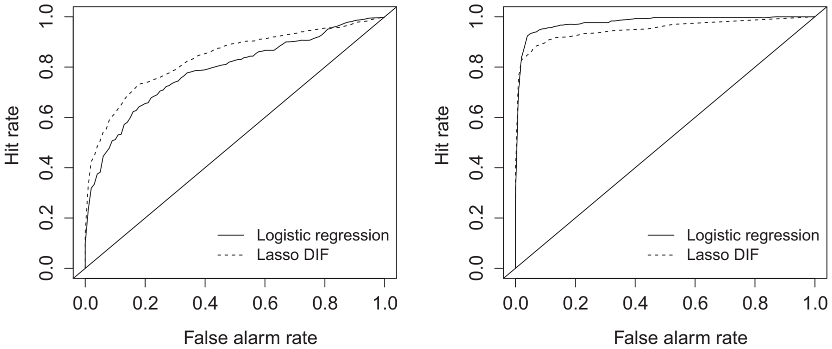

FIGURE 5. *Average ROC curves for logistic regression (LR) and LR lasso differential item functioning (DIF) methods in the cases where the differences in average receiver–operating characteristic (ROC) curve areas are maximal. Left panel: group sizes of 100 in both groups, presence of impact, 5% DIF items with DIF size 0.8 (areas: 0.286 for LR and 0.336 for LR lasso DIF). Right panel: group sizes of 1,000 in both groups, absence of impact, 5% DIF items with DIF size 0.4 (areas: 0.477 for LR and 0.450 for LR lasso DIF). Both cases deal with tests of 60 items.*

Tuerlinckx, De Boeck, & Magis, 2010; Magis & De Boeck, 2012; Magis & Facon, 2012; Penfield, 2001).

For the construction of the ROC curve approach (see, e.g., Hastie et al., 2009, pp. 314–317), the false alarm rates and hit rates were computed for increasing penalty parameters (for the LR lasso DIF method) and for decreasing α levels (for the LR and M-H methods). The ROC curves can then be plotted as step functions with false alarm rates on the x-axis and hit rates on the y-axis (examples of ROC curves are provided in Figure 5). For each generated data set within the simulation setting, one ROC curve is derived, and average ROC curves are computed for increasing false alarm rates of the three methods (lasso, LR, and M-H). Moreover, the area between the average ROC curve across replications and the identity line can be derived. These areas, further simply referred to as the *average areas*, are used as a global index of performance of the methods in the setting in question. The larger the area, the better the method is in discriminating between DIF and non-DIF items. The average area is thus a global comparative tool, sometimes referred to as the area under the curve (AUC; e.g., Perkins & Schisterman, 2006). However, it is limited to the settings with true DIF.

The average ROC areas can also be used to globally compare the six optimal penalty selection methods, in terms of both false alarm rates and hit rates. Each of the six methods relies on a particular λ value and can then be represented as a single point on the lasso ROC curve. The perpendicular distance between this point and the identity line is computed as a global measure of method accuracy. The larger the perpendicular distance, the better the tradeoff between false alarm and hit rates and the better the penalty selection method can discriminate between DIF and non-DIF items. The largest observable perpendicular distance is $\sqrt{2}/2 = 0.707$ and corresponds to a perfect discrimination between DIF and non-DIF items, yielding null false alarm rate and 100% hit rate. Note that some authors (such as Perkins & Schisterman, 2006) mention a slightly different summary statistic, referred to as the *Youden index J* (Youden, 1950), which is the maximum vertical distance from the ROC curve to the identity line.

Output Analysis

Three analyses were performed (a) a comparison of the LR lasso DIF approach with the LR and M-H methods, using the average ROC areas, (b) a comparison of the six methods for an optimal penalty parameter selection in the absence of DIF using only the false alarm rates, and (c) a comparison of the six methods in the presence of DIF using the perpendicular distances. For all three comparisons, the analysis of variance (ANOVA) approach, suggested by Magis and De Boeck (2012), was adopted.

First, because all statistics of interest (i.e., the average ROC curve area, the false alarm rate or the perpendicular distance) are in the unit interval, they are rescaled by applying a logit transformation. Second, an ANOVA is carried out with all design factors as covariates and all possible interactions included and with also the method as a factor (either lasso vs. LR or vs. M-H, or all optimal penalty selection methods). Then, the model is being reduced by removing first all nonsignificant terms (at the 5% significance level), then by removing all terms that explain less than 1% of the total variability. It is expected that some factors will be fully removed from the final model because they have no significant main effect or interaction effect or because they do not have a substantial effect size ($\geq$1%). The statistic can then be averaged out with respect to this factor. The final results will therefore be reported for a reduced number of factors, which will simplify the discussion.

Software

The full simulation study was performed with the R software (R Development Core Team (2012). The lasso penalization was run with the *glmnet* package (Friedman et al., 2010), while LR and M-H DIF analyses were performed with the *difR* package (Magis et al., 2010). The R code can be obtained from the first author. A sketch of the R code related to lasso penalization is provided in the Online Appendix.

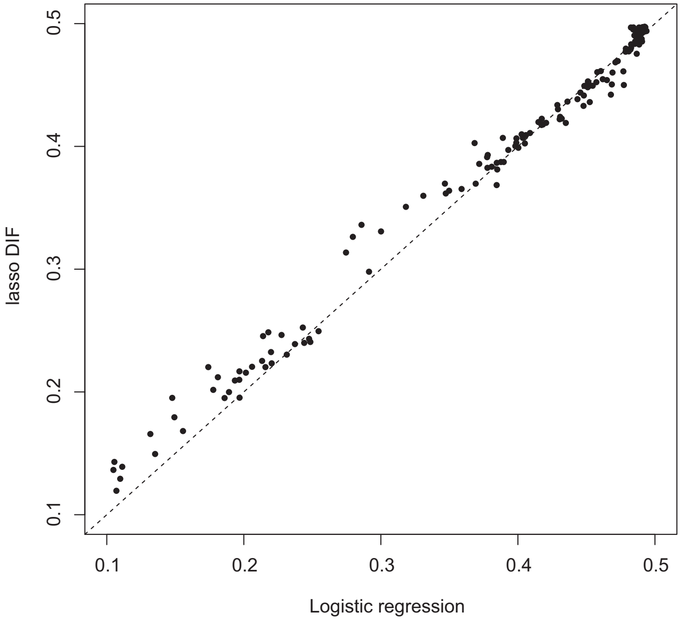

FIGURE 6. *Pairs of average receiver–operating characteristic (ROC) curve areas across all simulated settings with logistic regression (LR) and LR lasso differential item functioning methods.*

Results

First of all, the simulation study revealed very similar results between the LR and the M-H methods in terms of ROC curves. Only minor differences (up to the fourth decimal) occurred so that both methods are indistinguishable. For this reason, the forthcoming discussion focuses only on the LR method.

The global comparison of the LR and LR lasso DIF methods is discussed first. Then, the six methods for optimal penalty selection are being compared and discussed.

Standard Methods Versus LR Lasso DIF

Figure 5 displays two illustrative examples of average ROC curves in the two extreme situations that were observed among the 144 settings with DIF. In the left panel, the ROC curve of the LR lasso DIF approach always stands above the corresponding ROC curve for the LR approach, and the average areas are equal to 0.336 (for the LR lasso DIF) and 0.286 (for LR). It indicates that the former outperforms the latter across the whole range. In the right panel, however, the curves almost coincide up to hit rates of about 80% and then start to diverge slightly in favor of the LR method. The average areas equal 0.450 for the LR lasso DIF and 0.477 for LR.

Figure 6 displays the scatterplot for the average ROC curve areas of LR and LR lasso DIF for each of the 144 data sets with DIF (one point per setting).

This figure provides two insightful results. First, in most settings (98 of 144), the average ROC curve areas are larger for the LR lasso DIF than for the LR method. Interestingly, when the average areas of the LR method are smaller than 0.4, 53 settings out of 63 led to larger average areas for the LR lasso DIF method. This means that for the smaller areas and thus for the more difficult situations (to differentiate between DIF and non-DIF), the lasso method outperforms the LR method. For the larger areas and thus for the easier situations (easier to differentiate between DIF and non-DIF), the LR method performs equally well or better than the lasso method (in 38 out of 43 settings), but the maximum gap between the average areas is small. In other words, the LR lasso DIF approach performs better when the LR is not so efficient, while both methods tend to perform equally well (with a slight advantage to the LR) when both approaches are performing well in flagging DIF items.

The ANOVA revealed that, among the six design factors (test length, group size, presence/absence of item impact, percentage of DIF, DIF size, and DIF detection method), three are selected on the basis of significance and effect size: group sizes, DIF size, and impact, but the latter has a much smaller effect than the former two. The DIF detection method explains less than 1% of the variability but was nevertheless not removed from the final analysis because of our research objective. This model explains 96.1% of the total variance. Average ROC curve areas were consequently averaged out across the percentages of DIF and the test length. Figure 7 displays these average areas for increasing group sizes, in the absence (left panels) or presence (right panels) of item impact, and for DIF size 0.4 (top panels) and 0.8 (bottom panels).

Average areas increase with group sizes and with the DIF size and the increase with DIF size is even more important for smaller group sizes. Overall, the LR lasso DIF method returns larger or similar average areas than the LR approach, especially for small group sizes and in the presence of item impact. Conversely, the LR slightly outperforms the LR lasso DIF with larger group sizes, though both methods perform very similarly in terms of average ROC curve areas. These results are also in line with the finding from Figure 6 that the lasso method outperforms the LR method in the more difficult differentiation circumstances.

Different Lasso Methods: Performance in Absence of DIF

Let us now focus on the six optimal λ penalty selection methods only (CV3, CV5, CV10, AIC, BIC, and WIC). In the absence of DIF, the false alarm rates are best explained by two factors: the criterion and the test length and an interaction between these two factors. Neither the group size nor the impact

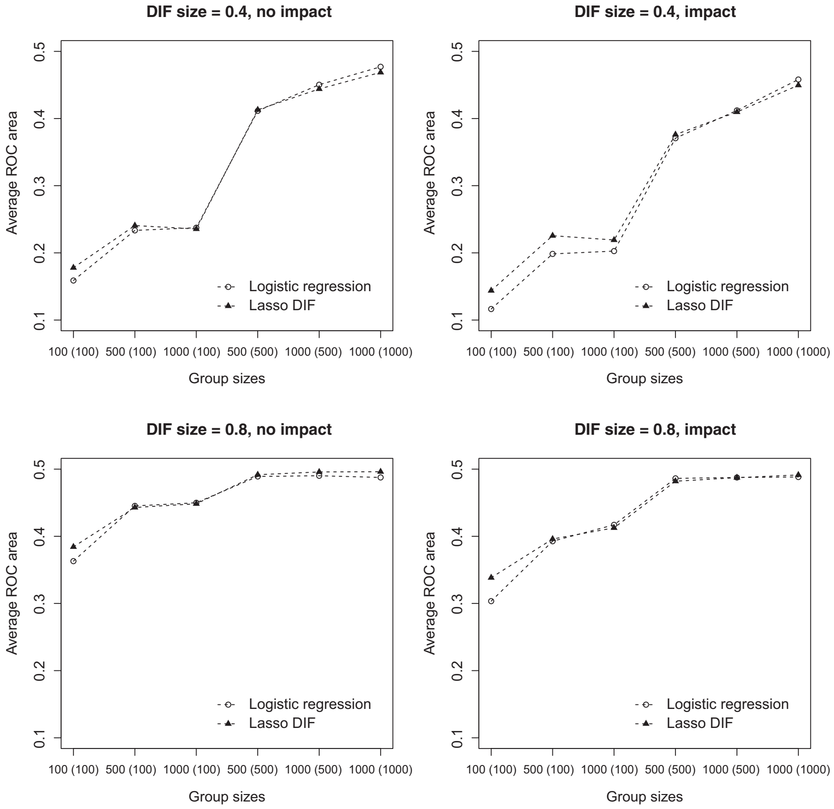

FIGURE 7. *Average receiver–operating characteristic (ROC) curve areas for the logistic regression (LR) and LR lasso differential item functioning (DIF) methods by group sizes, depending on absence (left panels) or presence (right panels) of item impact, and for DIF sizes 0.4 (top panels) and 0.8 (bottom panels).*

contributes enough to be retained as important explanatory factors. The selected model explains 98.4% of the total variance. The false alarm rates (averaged out across the group sizes and the presence/absence of impact) are displayed in Figure 8, per method and test length.

First, the false alarm rates tend to decrease with longer tests, which can be considered as an asset of the method. Further, the six criteria seem to differ, with the AIC criterion returning the largest rates and the BIC the smallest ones (almost 0). Roughly speaking, the criteria can be ordered (from larger to smaller rates) as follows: the AIC; the three CV methods, with decreasing Type I error rates as the number of folds decreases and with the WIC criterion in between the CV3 and CV5 methods; and the BIC criterion. The interaction between the methods and

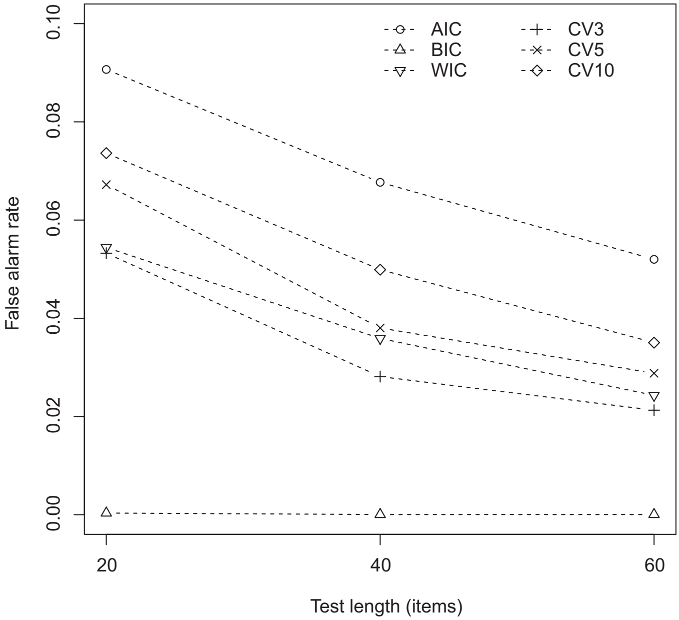

FIGURE 8. *False alarm rates in the absence of differential item functioning for the nine optimal penalty selection methods.*

the test length comes from the fact that the decrease in Type I error with the test length is not identical across the methods, as can be seen in Figure 8.

Different Lasso Methods: Performance in Presence of DIF

Finally, when DIF is present in the data, the perpendicular distances between the identity line and the average performance of the criterion (set by logit transformed average false alarm and hit rates) are used as summary statistics. They are best explained by four factors (in decreasing order of importance): criterion, group sizes, size of DIF, and percentage of DIF. The latter factor explains only 1.1% of the variability but was nevertheless retained. The final ANOVA model explains 90.1% of the total variance. Hence, perpendicular distances were averaged out across the presence/absence of item impact and the test length. They are displayed in Figure 9, per criterion and with increasing group size. The top and bottom panels refer to DIF sizes of 0.4 and 0.8, respectively, while left and right panels correspond to 5% and 10% of DIF, respectively.

First, there is a general trend of increased perpendicular distances when the group size increases and when the DIF size increases. For the latter, the increase is also more important for smaller group sizes than for the largest groups of

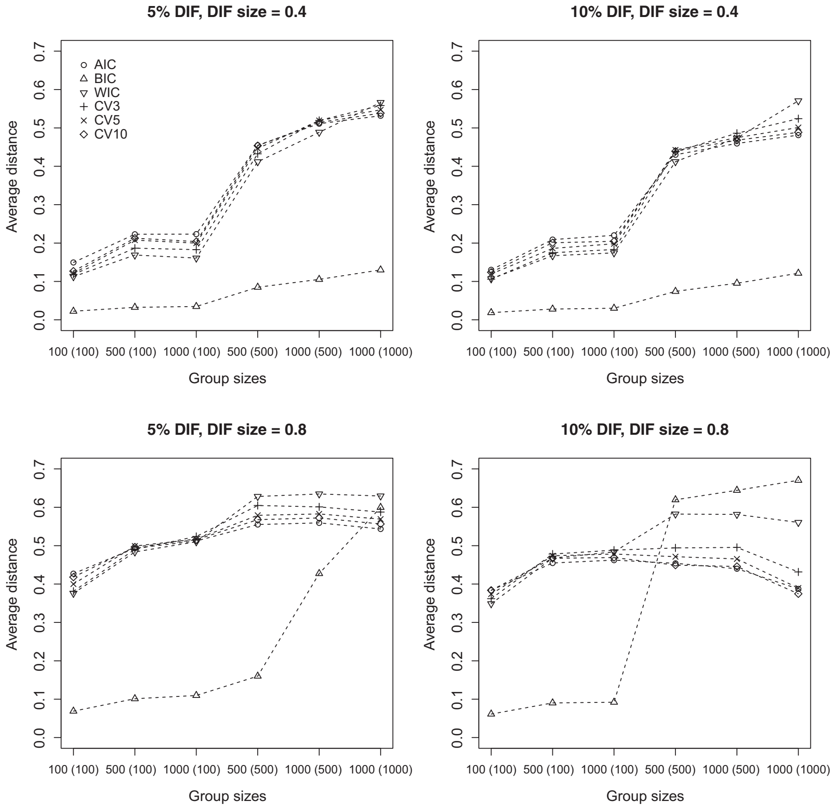

FIGURE 9. *Perpendicular distances of the nine optimal penalty selection criteria for increasing group sizes, with differential item functioning size 0.4 (left panels) and 0.8 (right panels) and for tests of 20 items (top panels), 40 items (middle panels), and 60 items (bottom panels).*

respondents. The effect of the percentage of DIF is less clear, in line with the ANOVA results. With small groups, the ordering of the six criteria is almost identical to what was observed in case of absence of DIF. Larger group sizes lead to different results, however. For instance, in the presence of large DIF size and large group sizes, the WIC criterion outperforms the standard ones, except for the BIC criterion when the percentage of DIF is large.

Globally, the WIC offers a good overall compromise. It performs best with large DIF sizes and large group sizes, but even for small groups its performance is reasonably good compared to the AIC and CV methods. The BIC criterion, on the other hand, is clearly not performing well as its perpendicular distances are most often far smaller than the other methods, except in a few specific cases.

Surprisingly, the BIC outperforms all methods (even the WIC) when group sizes and percentages of DIF and DIF sizes are large. The perpendicular distance reaches almost its maximum, whereas it decreases slightly for the other methods. This can mean that asymptotically the BIC is the better method and that the asymptotical result is approached sooner for rather clear cases (large percentage of large DIF), whereas other methods tend to overidentify DIF. However, this is the single case wherein one can observe such outperformance of the BIC criterion.

Discussion

The purpose of this article was to present a regularization approach to DIF. Starting from a test score–based logistic model that is widely used (Swaminathan & Rogers, 1990), a lasso penalization approach was developed to identify DIF items. Several optimal penalty selection methods were proposed, some based on usual information criteria, others developed specifically for this context. The simulation study revealed (a) the overall good performance of the LR lasso DIF approach compared with LR and M-H, (b) the relative inaccuracy of the usual information criteria such as AIC (in absence of DIF) and BIC (in both presence and absence of DIF), and (c) the relative overall good performance of the WIC criterion in the presence of DIF.

In this simulation study, emphasis was put on the LR but the M-H method returned very similar results, so that the two standard methods are undistinguishable in terms of performance. This is an obvious result since the data were generated under a one-parameter model and only uniform DIF effects were tested with the LR method. It was also found earlier that in such cases both methods are nearly identical in terms of DIF detection performance (Hidalgo & Lopez-Pina, 2004; Rogers & Swaminathan, 1993).

The main strength of this method stands in its flexibility. Lasso penalization can actually be applied to other types of methods. One may imagine, for instance, an IRT model equivalent to Equation 2 but with latent proficiency levels instead of the test scores. This is actually the guideline of Tutz and Schauberger's (in press) approach. The present approach, however, seems to work well, is computationally easy and does not require the IRT model estimation.

In this approach, the effect of the test score was constrained to be equal across items, as represented by Model 2. This assumption might actually be too strict as the relationship between an item score and the total test score might be item dependent, as in Model 1. This assumption, however, is not mandatory for fitting the model and performing DIF detection using lasso penalization. In a secondary simulation study (not shown here), this assumption was relaxed and DIF detection was performed by using exactly the same simulation design (but focusing on AIC, BIC, and WIC criteria only for sake of simplicity). The output of both this study and this additional analysis was then compared in terms of false alarm and hit rates. It was observed that though significantly improving the model fit, relaxing the assumption of equal test score effect across items had very limited impact on DIF results. All trends described in this analysis were preserved when the assumption was relaxed and very little differences (on average) were observed for the false alarm rates (less than 0.004) and the hit rates (less than 0.011). Though this should be further validated, it is concluded that allowing different test score effects across items improve the overall model fit but has very little impact on the efficiency in detecting DIF items with this lasso approach.

The method can easily be extended to more than two groups of respondents. It is straightforward to extend the definition of the DIF to any number of groups and to perform lasso penalization onto all DIF parameters for all groups simultaneously. One can then determine on the basis of the lasso approach which items function differently between which groups of respondents. The LR method has been extended to multiple groups' framework before (Magis, Raîche, Béland & Gérard, 2011; Magis & De Boeck, 2011), so that it can be used as a basis of comparison.

Another field of application for this approach is the identification of DIF in the presence of missing data. Standard DIF methods are very sensitive to missing data and their performance is affected in the absence of appropriate imputation methods (Finch, 2011; Robitzsch & Rupp, 2009). The lasso penalization approach might be a convenient approach to overcome this issue without requiring imputation.

Finally, the principle of penalization could even be considered slightly differently. For instance, Zou (2006) introduced an adaptive lasso penalization in which the penalty parameter λ may vary depending on the parameters. In other words, each penalized parameter α_{2j} can be penalized by a specific λ_j instead of a single overall penalty parameter λ. Since this method was shown to have some benefits, it may also be considered for the detection of DIF.

Acknowledgments

The authors wish to thank the editors and two anonymous referees for their helpful comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed the following financial support for the research, authorship, and/or publication of this article: This research was supported by a postdoctoral grant "Chargé de recherches" of the National Funds for Scientific Research (FNRS), Belgium, the IAP

Research Network P7/06 of the Belgian State (Belgian Science Policy), and the Research Fund of the KU Leuven, Belgium.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceeding of the second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370. doi:10.1007/BF02294361

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. In O. Opitz, B. Lausen, & R. Klar (Eds.), *Information and classification, concepts, methods and applications* (pp. 40–54). Berlin, Germany: Springer.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*, 253–260. doi:10.1177/014662168801200304

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, *6*, 269–279. doi:10.1207/s15324818ame0604_2

del Pino, G., San Martin, E., Gonzalez, J., & De Boeck, P. (2008). On the relationships between sum scorebased estimation and joint maximum likelihood estimation. *Psychometrika*, *73*, 145–151. doi: 10.1007/S11336-007-9023-2

DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, *70*, 961–972. doi:10.1177/0013164410366691

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368. doi:10.1111/j.1745-3984.1986.tb00255.x

Finch, W. H. (2011). The use of multiple imputation for missing data in uniform DIF analysis: Power and Type I error rates. *Applied Measurement in Education*, *24*, 281–301. doi:10.1080/08957347.2011.607054

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement*, *67*, 565–582. doi: 10.1177/0013164406296975

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement*, *47*, 432–457. doi: 10.1111/j.1745-3984.2010.00122.x

Friedman, J., Hastie, H., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Hidalgo, M. D., & Lopez-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. doi: 10.1177/0013164403261769

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.

Kim, J., & Oshima, T. C. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement*, *73*, 458–470. doi:10.1177/0013164412467033

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

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847–862. doi: 10.3758/BRM.42.3.847

Magis, D., & De Boeck, P. (2011). Identification of differential item functioning in multiple-group settings: A multivariate outlier detection approach. *Multivariate Behavioral Research*, *46*, 733–755. doi:10.1080/00273171.2011.606757

Magis, D., & De Boeck, P. (2012). A robust outlier approach to prevent Type I error inflation in DIF. *Educational and Psychological Measurement*, *72*, 291–311. doi: 10.1177/0013164411416975

Magis, D., & De Boeck, P. (2014). Type I error inflation in DIF identification with Mantel-Haenszel: An explanation and a solution. *Educational and Psychological Measurement*, *74*, 713–728. doi: 10.1177/0013164413516855

Magis, D., & Facon, B. (2012). Angoff's Delta method revisited: Improving the DIF detection under small samples. *British Journal of Mathematical and Statistical Psychology*, *65*, 302–321. doi: 10.1111/j.2044-8317.2011.02025.x

Magis, D., Raîche, G., Béland, S., & Gérard, P. (2011). A generalized logistic regression procedure to detect differential item functioning among multiple focal groups. *International Journal of Testing*, *11*, 365–386. doi: 10.1080/15305058.2011.602810

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel–Haenszel procedures. *Applied Measurement in Education*, *14*, 235–259. doi: 10.1207/S15324818AME1403_3

Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, *163*, 670–675. doi: 10.1093/aje/kwj063

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*, 421–425.

R Development Core Team (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502. doi:10.1007/BF02294403

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*, 197–207. doi:10.1177/014662169001400208

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

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement*, *69*, 18–34. doi:10.1177/0013164408318756

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. doi: 10.1177/014662169301700201

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. doi:10.1214/aos/1176344136

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, *58*, 159–194. doi:10.1007/BF02294572

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–70. doi:10.1111/j.1745-3984.1990.tb00754.x

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, *27*, 77–83. doi:10.3102/10769986027001077

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–170). Hillsdale, NJ: Erlbaum.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, *58*, 267–288.

Tutz, G., & Schauberger, G. (in press). A penalty approach to differential item functioning in Rasch models. *Psychometrika*. doi:10.1007/s11336-013-9377-6

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*, 113–144. doi: 10.1207/s15324818ame1702_2

Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods and Research*, *27*, 359–397. doi:10.1177/0049124199027003002

Wu, T.-J., Chen, P., & Yan, Y. (2013). The weighted average information criterion for multivariate regression model selection. *Signal Processing*, *93*, 49–55. doi:10.1016/j.sigpro.2012.06.017

Wu, T.-J., & Sepulveda, A. (1998). The weighted average information criterion for order selection in time series and regression models. *Statistics and probability Letters*, *39*, 1–10. doi:10.1016/S0167-7152(98)00003-0

Youden, W. J. (1950). An index for rating diagnostic tests. *Cancer*, *3*, 32–35.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society (series B)*, *68*, 49–67.

Zou, H. (2006). The adaptive lasso and its Oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429. doi:10.1198/016214506000000735

Authors

DAVID MAGIS is a research associate of the *Fonds de la Recherche Scientifique—FNRS* (Belgium), Department of Education, University of Liège, Boulevard du Rectorat 5, B-4000 Liège, Belgium, e-mail: david.magis@ulg.ac.be, and a research fellow of the KU Leuven, Belgium. His research interests include statistical applications in psychometrics and educational measurement. Please address all correspondence to this author.

FRANCIS TUERLINCKX is a professor of psychology at the KU Leuven (Belgium), Faculty of Psychology and Educational Sciences, KU Leuven, University of Leuven, Tiensestraat 102, 3000 Leuven, Belgium, e-mail: francis.tuerlinckx@ppw.kuleuven.be. His research interests include psychometrics, time series analysis, and mathematical modeling of human response times.

PAUL DE BOECK is a professor of quantitative psychology at the Ohio State University, 232 Lazenby Hall, 1827 Neil Avenue, Columbus, OH 43210, USA; e-mail: deboeck.2@osu.edu, and is also part-time affiliated with the KU Leuven, University of Leuven (Belgium). His research interests are item response theory and mixed models in general.

Manuscript received November 6, 2013
Revision received May 8, 2014; September 3, 2014
Accepted October 22, 2014