DIF Trees: Using Classification Trees to Detect Differential Item Functioning
Brandon K. Vaughn and Qiu Wang
Educational and Psychological Measurement 2010 70: 941 originally published online 27 August 2010
DOI: 10.1177/0013164410379326

The online version of this article can be found at:
http://epm.sagepub.com/content/70/6/941

Published by:
http://www.sagepublications.com

Additional services and information for Educational and Psychological Measurement can be found at:

Email Alerts: http://epm.sagepub.com/cgi/alerts
Subscriptions: http://epm.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav
Citations: http://epm.sagepub.com/content/70/6/941.refs.html

>> Version of Record - Dec 22, 2010
OnlineFirst Version of Record - Aug 27, 2010
What is This?
DIF Trees: Using Classification Trees to Detect Differential Item Functioning

Brandon K. Vaughn and Qiu Wang

Abstract

A nonparametric tree classification procedure is used to detect differential item functioning for items that are dichotomously scored. Classification trees are shown to be an alternative procedure to detect differential item functioning other than the use of traditional Mantel–Haenszel and logistic regression analysis. A nonparametric classification rule is examined through simulation and real data, and Type I error and power are compared with equivalent Mantel–Haenszel, logistic regression, and discriminant analyses.

Keywords
differential item functioning, classification, classification tree

Purpose

The purpose of this study was to present classification trees as a means of identifying differential item functioning (DIF) of items that are dichotomously scored. Classification trees are shown to be an alternative procedure to detect DIF other than the use of traditional Mantel–Haenszel (MH) and logistic regression (LR) procedures. A nonparametric classification rule is examined through simulation, and Type I error and power are compared with equivalent MH and LR.

Rationale and Background

Item bias represents a threat to the validity of test scores in many different disciplines. One characteristic of bias is DIF, in which examinees from different groups have
differing probabilities of success on an item after being matched on the ability of interest (Borsboom, Mellenbergh, & van der Linden, 2002).

The most popular procedure for DIF detection is the MH procedure. The MH procedure (Mantel, 1963; Mantel & Haenszel, 1959) is a three-way contingency table analysis to test for the dependence of two variables. Holland (1985) introduced the use of the MH method for the detection of DIF. The analysis examines 2 (subpopulations) × 2 (scoring categories) × k (categories or levels in the matching criterion) contingency tables. For this procedure, the matching criterion is not latent, but is the observed score.

A first stage of the MH method involves calculation of the common odds ratio (Mellenbergh, 1982). This estimate for item \( i \) (\( \hat{a}_{\text{MH}} \)) is obtained via

\[
\hat{a}_{\text{MH}} = \frac{\sum_{j=1}^{k} n_{1rj} n_{0fj} / n_{ij}}{\sum_{j=1}^{k} n_{0rj} n_{1fj} / n_{ij}},
\]

where \( j \) is the \( j \)th score level in the matching criterion (\( j = 1, \ldots, k \)). In Equation (1), \( n_{1rj} \) is defined as the number of subjects in the reference group who answered item \( i \) at score level \( j \) correctly, \( n_{1fj} \) is defined as the number of subjects in the focal group who answered item \( i \) at score level \( j \) correctly, and so on. The total numbers for the responses of subjects or group affiliation are indicated by a dot in the respective subscript. The grand total of subjects for item \( i \) at the \( j \)th scoring level is represented by \( n_{ij} \). This odds ratio can take on values in the range \([0, \infty)\), and a value of 1 indicates that no DIF is present. For \( \hat{a}_{\text{MH}} \) values lesser than 1, the studied item favors the focal group. The opposite would be true for \( \hat{a}_{\text{MH}} \) greater than 1. More specifically, \( \hat{a}_{\text{MH}} \) is the ratio of the odds that a subject from the reference group will get the item correct compared with the odds for a matched subject from the focal group.

The value for \( \hat{a}_{\text{MH}} \) is often transformed to \( \Delta_{\text{MH}} \) so that numerical values are easier to interpret, coinciding with research done at the Educational Testing Service. This transformation is

\[
\Delta_{\text{MH}} = -2.35 \ln(\hat{a}_{\text{MH}}).
\]

Roussos and Stout (1996a, 1996b) suggest the following guidelines to aid in the interpretation of DIF:

- Negligible or A-level DIF: Null hypothesis is retained or null hypothesis is rejected and \( |\Delta_{\text{MH}}| < 1 \).
- Moderate or B-level DIF: Null hypothesis is rejected and \( 1 \leq |\Delta_{\text{MH}}| < 1.5 \).
- Large or C-level DIF: Null hypothesis is rejected and \( |\Delta_{\text{MH}}| \geq 1.5 \).

Swaminathan and Rogers (1990) applied the LR procedure to DIF detection. The benefits of this procedure have been detailed in much research. They include (a) a procedure capable of detecting uniform and nonuniform DIF (unlike the MH procedure),
the capability to accommodate continuous and multiple ability estimates, and (c) comparable power in the detection of uniform and greater power in the detection of nonuniform DIF compared with the MH procedure (Li & Stout, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990).

DIF analysis can also be viewed from the perspective of classification procedures. Miller and Spray (1993) showed how a logistic discriminant function analysis could be used for DIF analysis on scored items. In this study, for item response $U$, observed test score $X$, and group indicator $G$ (e.g., reference/focal group), the discriminant function is

$$
\begin{align*}
P(G|X, U) &= \frac{e^{(1-G)(-z_0-x_1X-x_2U-x_3X+U)}}{1 + e^{(-z_0-x_1X-x_2U-x_3X+U)}},
\end{align*}
$$

One advantage to this technique is that $U$ can be either dichotomous or polytomous without requiring any special setup. The significance test for $\hat{z}_3$ provides a measure of nonuniform DIF, whereas $\hat{z}_2$ provides a measure of uniform DIF. Miller and Spray’s research indicated that this approach was simpler than LR in handling polytomous outcomes and was more powerful than a generalized MH procedure.

Each procedure has various assumptions that should be met for proper application of the analysis. For example, linear discriminant analysis (LDA) assumes multivariate normality of the classification variables and equal covariance matrices among the groups (Hair, Anderson, Tatham, & Black, 1988). Logistic regression is recommended in the cases where the multivariate normality assumption is not met (Tabachnick & Fidell, 2001). Both LDA and LR are not meant to handle analysis of complex non-linear data sets. Thus, there is need for nonparametric rules in classification. The MH procedure is perhaps the least stringent in terms of assumptions because it is nonparametric and has served psychometricians faithfully for years. Yet recent advances in nonparametric classification procedures suggest that an alternative to MH is possible that surpasses the strength of Miller and Spray’s (1993) technique and possibly the MH procedure as well.

The goal of classification trees is to predict or explain responses on a categorical dependent variable based on the measurements of one or more predictor variables. Tree structure rules are constructed by repeated splits of predictor variables into two or more subsets. The final subsets form a partition of the predictor variables. Vaughn and Wang (2008) have suggested that classification/regression trees (Breiman, Friedman, Olshen, & Stone, 1984) are more powerful (i.e., lower misclassification rates) than both LR and discriminant analysis in many situations. Under assumption violations, hit rates in classification were much better for regression trees. When parametric assumptions were met, all the techniques were comparable (with regression trees still having a slight advantage).

No research has been found that considers the use of classification trees for DIF detection. Thus, this study considers the issue of DIF analysis using this nonparametric tree-structured method (specifically the approach of the Classification and
Regression Trees procedure of Breiman et al., 1984). Comparisons are made with traditional DIF procedures in terms of Type I error and power.

**Classification Trees**

Classification trees are a nonparametric procedure that allows the use of both categorical and continuous variables in a type of discriminant analysis with the end result being a graph that looks like a tree. The construction of the tree is a type of variable selection with interactions between variables handled automatically. The tree attempts to find splits in the predictor variables that can be used to distinguish the classification outcome with high hit rates. Each split creates a subset (or node) of subjects based on one of the predictor variables. If the algorithm specifies more splits are needed for classification accuracy, this node is an intermediate node. If no further splits are needed for this path, it becomes a terminal node.

The classification tree can be thought to provide a probability model (Clark & Pregibon, 1992). For $k$ classes, each node $i$ of a classification tree has probability distribution $p_{ik}$. Because each case is assigned to a terminal node in the tree, each terminal node can be thought of as a random sample $n_{ik}$ from a multinomial distribution parameterized by $p_{ik}$. Venables and Ripley (2002) provide the deviance for the tree as

$$D = \sum_i D_i, \quad \text{where} \quad D_i = -2 \sum_k n_{ik} \log p_{ik}. \quad (4)$$

If we consider the benefit of splitting Node A into Nodes B and C, the reduction in deviance would be represented as

$$D_A - D_B - D_C = 2 \sum_k \left[ n_{Bk} \log \frac{p_{Bk}}{p_{Ak}} + n_{Ck} \log \frac{p_{Ck}}{p_{Ak}} \right]. \quad (5)$$

The probabilities in Equation (3) are unknown and are estimated from the proportions that would result from the split. In particular, the proportions would be estimated by

$$\hat{p}_{Bk} = \frac{n_{Bk}}{n_B}, \quad (6)$$

$$\hat{p}_{Ck} = \frac{n_{Ck}}{n_C}, \quad (7)$$

and

$$\hat{p}_{Ak} = \frac{n_B \hat{p}_{Bk} + n_C \hat{p}_{Ck}}{n_A} = \frac{n_{Ak}}{n_A}. \quad (8)$$

Equation (2) now becomes
A split is decided by the point that gives the maximum reduction in deviance. (This is done by an intensive computer search.) A terminal node is achieved when the number of cases reaching that node is small (e.g., \( n_i < 10 \)), or the node is considered to display a large degree of homogeneity. The procedure recommended by Breiman et al. (1984) is to measure the impurity of the distribution at a node and choose the split that reduces the average impurity the most. This measure is referred to as the Gini index:

\[
D_A - D_B - D_C = 2 \sum_k \left[ \frac{n_{Bk} \log \frac{n_{Bk} n_A}{n_{Ak} n_B}}{n_A n_B} + \frac{n_{Ck} \log \frac{n_{Ck} n_A}{n_{Ak} n_C}}{n_A n_C} \right].
\]  

(9)

Once again, because the probabilities are unknown, the node proportions are used as estimations.

**Data Source**

A simulation of item responses was set up for this study similar in design to Jodoin and Gierl (2001). Two simulation factors were considered in this study: reference/focal sample sizes, respectively (250/250; 500/250; 1,000/250; 500/500; 1,000/500; and 1,000/1,000), and two conditions of ability distributions (ref. and foc. \( N(0, 1) \) and ref. \( N(-0.5, 1) \)). Only one DIF pattern was considered in this present study (eight DIF items favoring the reference group). The simulated instrument consisted of 40 dichotomously scored items. The instrument length was not considered as a factor in this study. Three DIF values were established on a logit scale (.43, .64, and .86) and arbitrarily assigned to the simulated DIF items as detailed in Table 1. These DIF values represent three effects (small, medium, and large; see French, 2003; Paek, 2002; Vaughn, 2008). Only uniform DIF was considered in this study to make an adequate comparison with the MH procedure. An R program was used to simulate data from a one-parameter item response theory (1-PL IRT) model. The 1-PL model
was chosen so that a simple model could be compared among the procedures and that the total score would represent the subject totally. The generating parameters were established with difficulties of $-1$ (13 items), $0$ (14 items), and $1$ (13 items). The generating parameters for the focal group were the same with the exception of the adjustments for DIF items as detailed in Table 1. The number of replications for each data set was established at 1,000.

Method

The goal of classification trees is to predict or explain responses on a categorical dependent variable from their measurements on one or more predictor variables. The method for using classification trees to detect DIF will be similar to the approach taken for LR analysis (Swaminathan & Rogers, 1990). We examine the case of two-group classification ($i = 1, 2$; Loh & Shih, 1997). For DIF studies, the classification is chosen to be that of item response: correct versus incorrect response.

The measurement of $p$ predictor variables for group classification is possible and is notated by the $p$-dimensional vector $x = (x_1, \ldots, x_p)^T$. For a DIF study, the two main predictors are observable test score and group affiliation (reference group vs. focal group). Tree structure rules are constructed by repeated splits of these predictor variables into two or more subsets. The final subsets form a partition of the predictor variables. That is, classification trees produce subsets, $s$, of predictor variables that are associated with a group assignment, $r(s)$. All possible splits for each predictor variable at each subset are examined to find the split producing the largest improvement in goodness of fit. Only predictors that adequately predict group classification are selected in the final tree. For categorical predictor variables with $k$ levels present at a subset, there are $2^{(k - 1)} - 1$ possible contrasts between two sets of levels of the predictor. For ordered predictors with $k$ distinct levels present at a subset, there are $k - 1$ midpoints between distinct levels (Breiman et al., 1984).

An example DIF tree is shown in Figure 1. The two main predictors are observable test score (score) and group affiliation (reference/focal group). If the condition that is specified in the node is satisfied, one moves to the branch to the left. In this figure, subjects with scores greater than or equal to 17.5 are initially classified as correct response. However, for those subjects whose score is lower, classification depends on their group affiliation. For example, subjects who score greater than or equal to 13.5 (but less than 17.5) and who are in the reference group would be classified as correct response. However, subjects who are in the focal group would be classified as a wrong response. If the item were free of DIF, classification should not depend on group affiliation. The fact that it does depend on both group affiliation and observed score gives evidence of DIF for this item.

The tree method for detecting DIF in dichotomous outcomes was compared with the traditional MH procedure (Mantel & Haenszel, 1959) and the LR approach with observable matching criterion (Swaminathan & Rogers, 1990). All procedures were done in the statistical program R.
Identification of DIF for each traditional procedure was carried out using established criteria as documented by Rogers and Swaminathan (1993) and Vaughn (2006, 2008). For the tree DIF procedure, DIF occurred based on the “improvement” it made on the overall model. Improvement can be defined as the improvement in deviance that a variable has in the overall model classification. If the “group affiliation” variable made a large improvement in model classification, given the score variable in the model, then this was deemed as an indication of DIF. Two criteria were considered: the maximum improvement for group affiliation and the total sum improvement for group affiliation. The value used for each was calculated using values from an initial simulation for each condition. These values changed for each simulation condition, which seems to be a factor of sample size and ability distribution. Further studies are needed to find a formulated approach to the choice of values in a given situation. Roughly, it appears that a good value for the summed improvement is focal group size divided by 100, with the maximum improvement one or two points less. Type I error and power were assessed for all simulation runs. Only uniform DIF

![Figure 1. Sample differential item functioning tree](image-url)
was considered in this study, although a follow-up study is being conducted to consider the effect on nonuniform DIF.

In the context of DIF, a Type I error is defined as the incorrect classification of an item as displaying DIF when, in fact, it does not. As equally important is the consideration of power, which is defined as the correct classification of an item as displaying DIF when, in fact, it does. Both were considered in this study. The importance of each is an important consideration of the technique (although in some applications, one might be of more importance to a psychometrician).

Results

The average Type I errors (non-DIF items) and the average power (DIF items) for the simulation conditions are given in Tables 2 and 3. These results indicate the feasibility in using classification trees to detect DIF. Based on these results, the following is apparent:

1. The tree method has similar Type I error rates as compared with MH and LR procedures for equal distributional conditions. In the case that the focal group is centered at a lower ability level, the tree procedure at times had lower Type I errors.

Table 2. Simulation Results for Eight DIF Items: Ref. and Foc. N(0, 1)

<table>
<thead>
<tr>
<th>Sample Size Condition</th>
<th>DIF Rule</th>
<th>Type I Error*</th>
<th>Overall Power*</th>
<th>Power (Low DIF Effects)*</th>
<th>Power (Medium DIF Effects)*</th>
<th>Power (High DIF Effects)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>N_R = 250; N_F = 250</td>
<td>MH</td>
<td>0.09</td>
<td>0.60</td>
<td>0.31</td>
<td>0.68</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.09</td>
<td>0.61</td>
<td>0.31</td>
<td>0.69</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Tree</td>
<td>0.10</td>
<td>0.64</td>
<td>0.37</td>
<td>0.70</td>
<td>0.94</td>
</tr>
<tr>
<td>N_R = 500; N_F = 250</td>
<td>MH</td>
<td>0.06</td>
<td>0.61</td>
<td>0.28</td>
<td>0.70</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.05</td>
<td>0.61</td>
<td>0.27</td>
<td>0.70</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Tree</td>
<td>0.06</td>
<td>0.66</td>
<td>0.36</td>
<td>0.73</td>
<td>0.98</td>
</tr>
<tr>
<td>N_R = 1,000; N_F = 250</td>
<td>MH</td>
<td>0.04</td>
<td>0.62</td>
<td>0.26</td>
<td>0.73</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.04</td>
<td>0.62</td>
<td>0.25</td>
<td>0.74</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Tree</td>
<td>0.04</td>
<td>0.72</td>
<td>0.43</td>
<td>0.83</td>
<td>0.99</td>
</tr>
<tr>
<td>N_R = 500; N_F = 500</td>
<td>MH</td>
<td>0.02</td>
<td>0.62</td>
<td>0.24</td>
<td>0.76</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.02</td>
<td>0.62</td>
<td>0.24</td>
<td>0.76</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Tree</td>
<td>0.02</td>
<td>0.71</td>
<td>0.40</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td>N_R = 1,000; N_F = 500</td>
<td>MH</td>
<td>0.01</td>
<td>0.61</td>
<td>0.20</td>
<td>0.78</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.01</td>
<td>0.61</td>
<td>0.20</td>
<td>0.78</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Tree</td>
<td>0.01</td>
<td>0.74</td>
<td>0.43</td>
<td>0.89</td>
<td>1.00</td>
</tr>
<tr>
<td>N_R = 1,000; N_F = 1,000</td>
<td>MH</td>
<td>0.00</td>
<td>0.62</td>
<td>0.16</td>
<td>0.84</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.00</td>
<td>0.62</td>
<td>0.15</td>
<td>0.84</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Tree</td>
<td>0.00</td>
<td>0.79</td>
<td>0.49</td>
<td>0.95</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: DIF = differential item functioning; MH = Mantel–Haenszel; LR = logistic regression.

a. Average value across related items of all replications.

Downloaded from epm.sagepub.com at THE HONG KONG INSTITUTE OF ED on March 18, 2012
2. In almost all situations, the tree procedure had greater overall power in detecting actual DIF. Thus, in reference to an instrument containing a variety of DIF effects, the tree procedure appears to have a higher success rate at actual DIF detection. This tendency improves as the amount of data increases, which is not always the case with MH and LR.

3. In terms of low and medium DIF effects, the tree has tremendous detection rates as compared with MH and LR. This is important for researchers who might need another useful DIF detection technique for items that may display smaller DIF effects. Because very large DIF effects are probably rare in actual test development situations, this result suggests that a tree approach is useful. The power became worse for MH and LR as the sample size increased, possibly because of more “noise” that makes the detection of smaller DIF effects harder. However, this was not true of the tree procedure.

4. In terms of high DIF effects, all techniques were equally good. This result is not surprising, yet it is important to establish that the tree procedure is functioning properly in this situation.

Although these results do not find a definitive solution to detecting DIF for smaller sample sizes, they do provide some evidence that a tree procedure would be a welcome addition in such situations. And in the case of typical sample sizes, the tree procedure does appear to be a great method to consider in the detection of small to medium DIF.

### Table 3. Simulation Results for Eight DIF Items: Ref. \( N(0, 1) \), Foc. \( N(-0.5, 1) \)

<table>
<thead>
<tr>
<th>Sample Size Condition</th>
<th>DIF Rule</th>
<th>Type I Error(^a)</th>
<th>Overall Power(^a)</th>
<th>Power (Low DIF Effects)(^a)</th>
<th>Power (Medium DIF Effects)(^a)</th>
<th>Power (High DIF Effects)(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_R = 250; N_F = 250 )</td>
<td>MH</td>
<td>0.11</td>
<td>0.61</td>
<td>0.32</td>
<td>0.67</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.10</td>
<td>0.60</td>
<td>0.30</td>
<td>0.68</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Tree</td>
<td>0.03</td>
<td>0.68</td>
<td>0.44</td>
<td>0.72</td>
<td>0.98</td>
</tr>
<tr>
<td>( N_R = 500; N_F = 250 )</td>
<td>MH</td>
<td>0.07</td>
<td>0.61</td>
<td>0.29</td>
<td>0.70</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.06</td>
<td>0.62</td>
<td>0.29</td>
<td>0.70</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Tree</td>
<td>0.04</td>
<td>0.77</td>
<td>0.57</td>
<td>0.81</td>
<td>1.00</td>
</tr>
<tr>
<td>( N_R = 1,000; N_F = 250 )</td>
<td>MH</td>
<td>0.05</td>
<td>0.61</td>
<td>0.28</td>
<td>0.71</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.04</td>
<td>0.62</td>
<td>0.28</td>
<td>0.72</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Tree</td>
<td>0.02</td>
<td>0.75</td>
<td>0.53</td>
<td>0.80</td>
<td>1.00</td>
</tr>
<tr>
<td>( N_R = 500; N_F = 500 )</td>
<td>MH</td>
<td>0.03</td>
<td>0.62</td>
<td>0.26</td>
<td>0.74</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.03</td>
<td>0.62</td>
<td>0.26</td>
<td>0.74</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Tree</td>
<td>0.00</td>
<td>0.69</td>
<td>0.42</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>( N_R = 1,000; N_F = 500 )</td>
<td>MH</td>
<td>0.02</td>
<td>0.62</td>
<td>0.21</td>
<td>0.78</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.01</td>
<td>0.62</td>
<td>0.21</td>
<td>0.78</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Tree</td>
<td>0.02</td>
<td>0.85</td>
<td>0.69</td>
<td>0.91</td>
<td>1.00</td>
</tr>
<tr>
<td>( N_R = 1,000; N_F = 1,000 )</td>
<td>MH</td>
<td>0.00</td>
<td>0.62</td>
<td>0.16</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.00</td>
<td>0.62</td>
<td>0.16</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Tree</td>
<td>0.01</td>
<td>0.88</td>
<td>0.73</td>
<td>0.94</td>
<td>1.00</td>
</tr>
</tbody>
</table>

\(^a\) Average value across related items of all replications.

---

Note. DIF = differential item functioning; MH = Mantel–Haenszel; LR = logistic regression.
effects as compared with MH and LR procedures. However, more studies need to be considered at this point (as detailed in the Method section) before suggesting that the tree procedure is a superior technique over MH and LR in those situations.

**Implications**

The DIF tree approach differs from traditional MH and logistic regression DIF assessments in that the classification of group membership (reference/focal group) is central. If the group affiliation ends up in the final subset of predictors, this procedure is indicating that given observed ability score, item response is affected by the reference/focal group membership (thus evidence of DIF). Like the discriminant analysis approach (Miller & Spray, 1993), the tree method has advantages over LR and MH procedures in the ease of handling polytomous outcomes (though not considered in this study). The same DIF tree procedure would be used in both types of outcomes, which would be useful in mixed format instruments. The technique is easily expanded to consider differential test functioning as well.

Unlike discriminant analysis, tree results give immediate information regarding the details of DIF. That is, if DIF occurs, one can immediately see at which ability levels this DIF is occurring based on the split of the tree. Whether DIF favors the focal or reference group is also easily seen based on the item response split rule of the final tree. A researcher would know this information immediately without need for further analysis or calculations. Furthermore, the DIF tree method is invariant to assumption violations of parametric procedures (similar to the MH procedure). Thus, the potential worth of the DIF tree analysis is a procedure that combines the strengths and simplicity of an MH procedure with a more powerful DIF detection.

**Limitations and Future Study**

There are numerous limitations to this study that are currently being considered: skewed ability distributions, varying amounts of DIF items (including more of a mixture of DIF items favoring both focal and reference groups), and real data application. In addition, we are currently considering a mathematical derivation of the max/sum criteria value for optimal classification and identification of DIF. Also, extending the rule to detect nonuniform DIF is important and should be considered as well (and is in our current research), as is the performance of this procedure with mixed formats (dichotomous and polytomous responses). Last, we considered more of an effect size approach in detecting DIF for traditional procedures in this study, and the current tree procedure lacks it. Future research should focus on the derivation of an effect size for DIF analysis using the tree procedure. We would also like to look at incorporating significance tests for both MH and LR, as well as $\Delta R^2$ for LR when comparing it with the tree procedure. Because the tree procedure does appear to be a more powerful technique at detecting low to medium DIF effects in this study, it should be more heavily scrutinized under these varying conditions.
Declaration of Conflicting Interests

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

References


