Logistic Regression and Model Based Recursive Partitioning for Item Evaluation

W. Holmes Finch
Ball State University

Brian F. French
Washington State University

Complex sampling plans are common in large datasets that are developed in national and international contexts. The multilevel structure of such data is handled most frequently by multilevel models. However, analysis becomes complex under such conditions, particularly if estimates (e.g., regression slopes) of individual units is desired. Such cases include validity studies where differential item function analyses (DIF) are conducted. This simulation study evaluated the accuracy of model based recursive partitioning (MBRP) using logistic regression for DIF detection under various conditions. Results suggest that MBRP is a promising technique for linear and logistic models, including those used for DIF detection. Implications and future directions for research are discussed.

Increasingly, social science researchers are making use of large scale datasets based on complex sampling plans. Persons, for example, are sampled from, and therefore nested within, larger organizing units such as states in the case of the National Assessment of Educational Progress (NAEP) and nations in the cases of the Program for International Student Assessment (PISA) and the Progress in International Reading Literacy Study (PIRLS). These data are then used to examine relationships among variables such as academic performance, demographic characteristics, and psychological measures of the individuals in the sample (Organization for Economic Co-operation and Development, 2013; 2014). When researchers use such data to fit a specific model, such as a linear regression with achievement as the dependent variable and student socioeconomic status as an independent variable, they make a tacit assumption that the model parameters are the same or invariant across the organizing units. Typically, data analysis in these cases would involve the use of multilevel models to account for the nested structure of the data (Raudenbush & Bryk, 2002). These models can allow for random coefficients terms such that different clusters in the nested structure are allowed to have different model parameters. However, cluster-specific model parameter estimates (e.g., slopes for each nation) are not easy to obtain from such models. In addition, although the multilevel model results would indicate whether there are differences in model parameters across the organizing units (e.g., nations), there would be a lack of certainty as to which coefficients differed for which nations. In other words, the researcher would know that a given parameter was not consistent across nations, but he or she would not know whether the parameter was equal for some nations and not for others.

In this simulation study, the performance of an alternative methodology that can be used with complex data, model based recursive partitioning (MBRP), was examined. MBRP is related to recursive partitioning models such as classification and regression trees (CART) and random forest. With MBRP; however, rather than partitioning the sample based on values of the variables, partitioning is based on differences in model parameter estimates. The ability of MBRP to correctly divide the sample based on the higher-level organizing units was the focus of this research. In this study, we (a). review the context in which such modeling might be useful, (b). review the basic principles underlying MBRP, (c). describe the study methodology, and (d). present results and discuss implications.

CART

MBRP is based upon the CART methodology described by Breiman, Friedman, Stone, and Olshen (1984). With CART, a prediction model, including one dependent and several independent predictor variables, is specified. Consider, for example, a categorical outcome variable (e.g., a student is proficient at reading or not proficient at reading) with a mix of categorical (e.g., gender) and continuous (e.g., family income) predictors. CART begins with all members of the sample in a single grouping, which is referred to as node 1. The algorithm then searches the entire sample for the binary partition among the predictors, which results in the most homogeneous split possible, in terms of the outcome variable proficiency. This split creates child nodes and individuals are placed into the appropriate node based upon their predictor variable value. For example, if the optimal split is on gender, then males are placed into one child node (e.g., node 2) and females are placed into the other child node (e.g., node 3). The CART algorithm next investigates potential splits in each of these child nodes. Again, with the goal of creating the most homogeneous child nodes from each with respect to the outcome variable. Continuing with the example, assume that the optimal split for node 2 is on family income at a value of $50,000; such that those with family income less than $50,000 are placed in child node 4 and those with family income of $50,000 or higher are placed in child node 5. This partitioning continues until further separation does not yield increased homogeneity in the dependent variable. At this point, the tree stops growing and the final child nodes (i.e., the terminal nodes) are referred to by the researcher in order to understand how members of the sample are differentiated based upon their gender and family income, with respect to reading proficiency.

MBRP

MBRP uses the same partitioning methodology as CART (Brieman et al., 1984; Kim & Loh, 2001; Zeileis, Hothorn, & Hornik, 2008). However, whereas CART builds a tree that maximizes partitioning group differences on the mean of the dependent variable, MBRP maximizes partitioning group differences based on the parameter estimates of a statistical model such as regression. In the context of this study with large datasets, the partitioning groups might be nations represented in such programs as PIRLS, PISA, or U.S. states as with NAEP. As described in Zeileis et al., this recursive partitioning approach uses the following steps:

1. Fit the model of choice (e.g., logistic regression [LR]) to all observations in the current node (e.g., node 1).
2. Assess parameter instability (defined below) for each independent variable in the model and select the one with the highest instability. If no instability is present, the algorithm stops. If instability is present, proceed to step 3.
3. Compute the split point of the model parameter value for the variable identified in Step 2 that optimizes partitioning group (e.g., nations) separation by yielding the most homogeneous (with respect to model parameters) child nodes possible.
4. Divide the node based upon this split point to create two child nodes in which members of the partitioning variable, with the most similar parameter estimates from step 3, are placed together.
5. Repeat steps 1 through 4 until stability is achieved for all of the model parameters.

The resulting tree consists of a set of terminal nodes for each of which model parameter estimates are obtained. These estimates are examined to determine how the terminal nodes differ from one another, with respect to the relationships between the independent and dependent variables, in the case of the statistical model employed (e.g., regression) for the analysis.

A key aspect of the MBRP algorithm is the determination of parameter stability. The stability assessment is made using the test statistic $\lambda_{\chi^2}(W_j)$, which takes the form:

$$\lambda_{\chi^2}(W_j) = \sum_{c=1}^{C} \frac{|I_c|^{-1}}{n} \left\| \Delta_{I_c} W_j \left(\frac{i}{n}\right) \right\|_2^2 \qquad (1)$$

In (1), $\Delta_{I_c} W_j$ measures whether there are systematic fluctuations in the score function associated with the regression model that is fit to the data, across categories (C) of the partitioning variable (e.g. nations). The value I_c is the number of individuals in category c of the partitioning variable (e.g., individuals from the United States), I is the total number of individuals in the node, and n is the total sample size. If the model parameters are comparable for all levels of C, the score function should fluctuate randomly around its mean of 0. On the other hand, systematic fluctuations in the score function from one level of C to another would indicate instability in the model parameters, such as regression coefficients, meaning that different model forms are needed for different levels of the partitioning variable. Calculation of the statistic in (1) for a given node requires that the regression model be fit once for all individuals in the node, after which the score function values are reordered and aggregated to calculate $\lambda_{\chi^2}(W_j)$ for each possible combination of the C levels of the partitioning variable. Thus, as an example, a separate regression model is estimated for every possible combination of the partitioning group categories (e.g., nations) and the score function for each such grouping is retained. The parameter $\lambda_{\chi^2}(W_j)$ is asymptotically distributed as $\chi^2$with $k * (C - 1)$ degrees of freedom, where k is the number of model coefficients. A statistically significant $\lambda_{\chi^2}(W_j)$ value for a node indicates that the model parameters are not equal across the levels of C in the node (i.e., are unstable) in which case, a binary split is made based on the partitioning variable. To determine where the split will be made, an exhaustive search of all possible combinations of the partitioning variable (i.e., the Nation variable in this example) is made where for each partition, the model in (1) is fit. For each of these partitions, the sum of the score function is calculated across individuals within the nodes and the split that minimizes the score function sum across the two resulting child nodes is selected. These steps are repeated for each of the resultant child nodes. Partitioning stops when parameter stability has been reached (i.e., when $\lambda_{\chi^2}(W_j)$ is not statistically significant) indicating that within the existing nodes, the model parameters are consistent across levels of the partitioning variable.

Example of MBRP

As a simple example of how MBRP works in practice, consider the following scenario. A researcher using PIRLS data is interested in determining whether there is uniform differential item functioning (DIF) for items on a reading assessment. There are a total of 43,555 examinees from 11 nations who completed the reading test. Of particular interest is whether DIF exists with regard to the language spoken by the mother in the home (i.e., 0 = Not the language of the test, 1 = Language of the test) and a continuous score representing the available educational resources in the home, where a higher score indicates more such resources. The dependent variables in each of the LR models are the responses to the reading assessment items (i.e., 0 = Incorrect and 1 = Correct). The LR model for this uniform DIF assessment, thus, takes the form:

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 score + \beta_2 language + \beta_3 resources \qquad (2)$$

The independent variables in (2) are the total score on the assessment: mother's language spoken in the home and available educational resources. The term $\pi(x)$ is the probability of a correct item response. The partitioning variable is nation, which contains 11 categories.

The MBRP analysis for this problem identified terminal nodes for which the model parameters β_0, β_1, β_2, and β_3 were homogeneous based on the test in equation (1). The resulting MBRP tree for item 1 appears in Figure 1. Three terminal nodes were extracted by the algorithm, with nations 372, 380, 440, 528, 616, 620, and 643 constituting the first node; nations 578, 642, and 703 constituting the second node; and nation 752 being in the third node. Below each terminal node are three graphs showing the relationships between the independent variables and the outcome, item response. The first graph demonstrates the relationship between total test score on the *x*-axis and the probability of a correct item response on the *y*-axis. Thus, we can see that for each node, the higher the test score, the greater the probability that an individual will respond to the item correctly. Similarly, the second graph manifests the relationship between mother's language in the home and the probability of a correct item response. Finally, the third graph shows the relationship between educational resources available in the home and a probability of a correct item response. These relationships can perhaps more clearly be seen in Table 1, which includes the coefficients for each variable from the LR model. As was apparent in Figure 1, for each terminal node, the relationship between total test score and probability of a correct item response was positive. With respect to Mother's language, terminal node 2 had a statistically significant negative coefficient, which indicates that examinees whose mothers speak a language in the home other than that of the test had a lower probability of answering the item correctly. This result indicates the presence of uniform DIF for the nations in this node. For nodes 1 and 3, this relationship was not statistically significant. For none of the terminal nodes, was there a statistically significant relationship between educational resources in the home and the item response. In summary, there was evidence of uniform DIF with respect to mother's language in the home for the three nations in terminal node 2: The Netherlands, Romania, and the Slovak Republic. For none of the other nations, was uniform DIF present. Furthermore, there was no DIF associated with the educational resources in the home.

Study Goals

The goals of this simulation study were to examine the performance of the MBRP algorithm in terms of its ability to recover the appropriate number of partitioning categories based on differences in LR model parameters. Specifically, coefficient values for one variable in a LR model were simulated to differ among partitioning subgroups and the number of terminal nodes, and parameter estimate differences among the terminal nodes were recorded in order to evaluate the performance of MBRP. Several factors were manipulated in the study design to examine influences on the performance of the algorithm. Specific hypotheses, with respect to some of these manipulated factors, are included in the description of the study methodology.

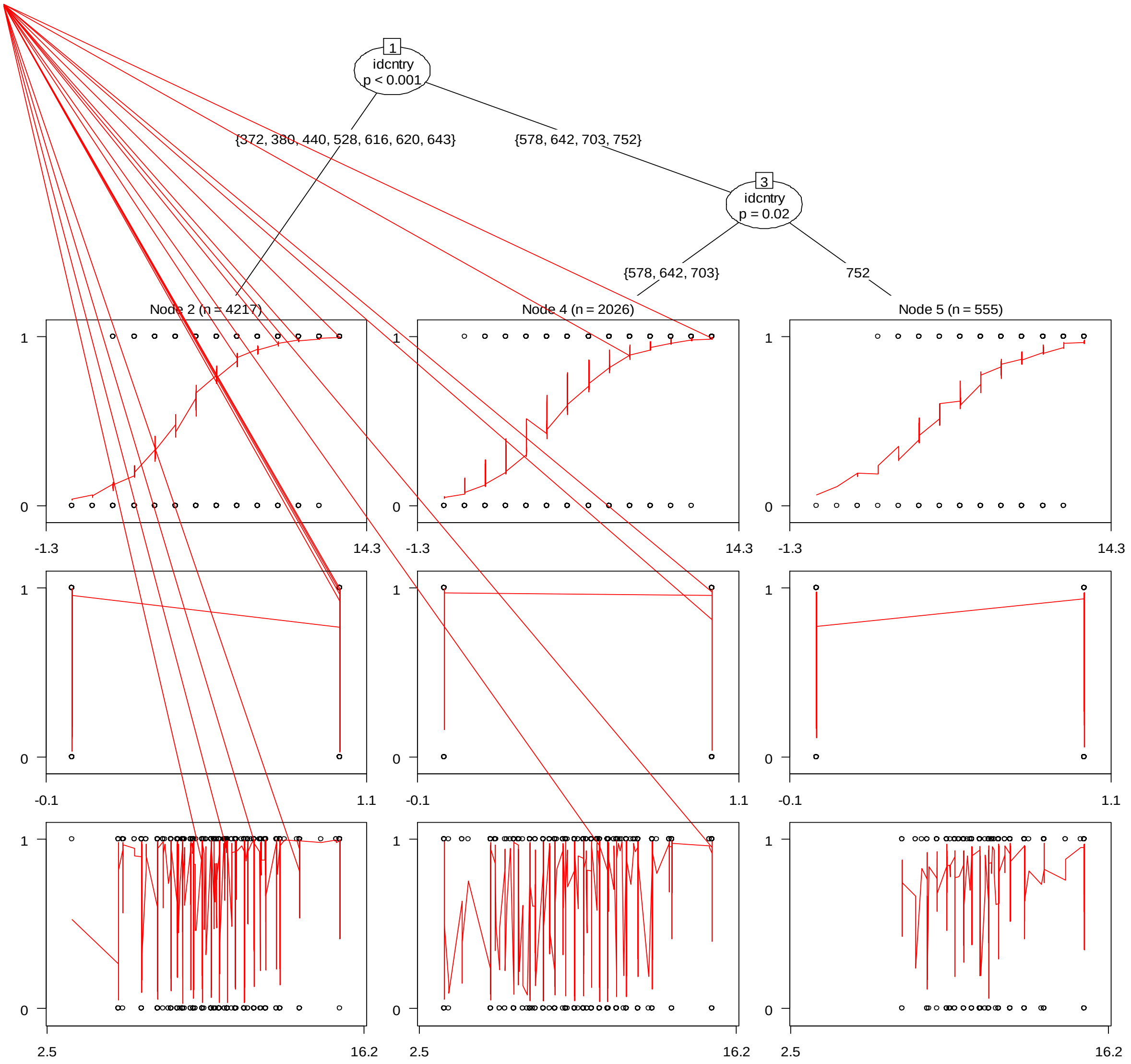

Figure 1. MBRP for item 1 on the PIRLS reading assessment.

Table 1. DIF Results Based on MBRP Tree for Item 1.

Item	Nations[1]	Total score	Mother's language	Educational resources
1	372, 380, 440, 528, 616, 620, 643	0.64*	0.06	0.07
	578, 642, 703	0.57*	**-0.73***[2]	-0.04
	752	0.45*	-0.41	0.07

[1] 372=Ireland, 380=Italy, 440=Lithuania, 528=Netherlands, 578=Norway, 616 =Poland, 620=Portugal, 642=Romania, 643=Russian Federation, 703=Slovak Republic, 752=Sweden

*Indicates statistically significant coefficient (α=0.05)

[2] Bold indicates the presence of uniform differential item functioning

Method

To investigate the performance of MBRP based on LR for identifying DIF, a simulation study was used. For each combination of conditions described below, 1,000 replications were conducted. The data generating LR model for a dichotomous dependent variable included one continuous (*cont*) and one categorical (*cat*) independent variable and an interaction between these variables. In the context of DIF assessment, *cont* would correspond to the total test score and *cat* would correspond to the grouping variable for which uniform DIF is being assessed. The interaction of the two variables allows for testing for the presence of non-uniform DIF. The model was defined as:

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 cont + \beta_2 cat + \beta_3(cont * cat) \quad (3)$$

The following factors were manipulated in the simulation study in order to assess the performance of the MBRP algorithm under various conditions.

Number of Partitioning Variable Groups

A total of 2 and 10 partitioning variable groups were simulated. These values were selected in order to reflect a relatively small number, as well as a large number, of partition groups. These conditions are similar to analyses conducted across nations, in PIRLS or PISA datasets or other such large scale international assessment programs, to smaller DIF investigations with two groups.

Partitioning Group Size and Group Size Ratio

For one set of conditions, the partitioning variable groups were all of equal size with each consisting of 250, 500, or 1,000 individuals. In the other set of conditions, the partitioning variable groups were of unequal size with half of the groups being half of the size of the other groups. Thus, for example, in the 10 partition groups 1,000 sample size condition, 5 of the groups included 1,000 individuals and the other 5 contained 500. These conditions allow for examination of how various ratios of sample size will influence the accuracy of the results.

Model Coefficient Values

Values of β_2 in model (3) were simulated to be 0, 0.2, 0.4, 0.6, and 0.8. This reflected no differences in item responses to large differences in items responses across groups. Values of β_0 were simulated to be 1 across all conditions, β_1 was simulated to be 0.5, and β_3 was simulated to be 0. In the context of DIF assessment, the data generating model was designed to simulate uniform DIF such that there were group differences in terms of the dichotomous outcome variable value (i.e., item response) after conditioning on the continuous predictor variable (i.e., cont or test score).

Differences in β_2 across Partitioning Groups (Partition DIF)

The β_2 parameter was simulated to be both the same and different across levels of the partitioning variable. When equal, all partitioning groups had the same coefficient values for model (3). This condition allowed for the assessment of MBRP when no partitioning was actually necessary. In the second condition, β_2 in model (3) differed among the partitioning groups by 0.2, 0.4, and 0.8. In other words, the impact of cat on the dependent variable was simulated to differ for the different partitioning groups. In this condition, each partitioning group had a unique value of β_2 that differed by the said amounts. As an example, in the 2 partitioning groups 0.2 difference case in which β_2 was simulated at 0.4, group 1 had a value of 0.4 and group 2 had a value of 0.6. In the 10 partitioning groups 0.2 difference with the baseline β_2 value of 0.4, the partitioning groups' β_2 values were as follows: Group 1 = 0.4, Group 2 = 0.6, Group 3 = 0.8, …, Group 10 = 2.2. Similar patterns of partition group β_2 differences were used for the other conditions.

Impact

The 2 levels of the categorical predictor variable in model (3) were simulated to have either the same mean on the continuous independent variable, or to have means that differed by 0.5 on this variable. The standard deviations did not differ and were constrained to 1.0. The differences in ability across groups generally inflates Type I error in investigations of DIF. Thus, we expected accuracy to be hindered in the presence of impact compared to the absence of impact.

Simulation Outcome Variables

The outcome variables for the simulation were the number of terminal nodes recovered by the MBRP algorithm, and the root mean square (RMS) of the LR parameter estimates across the terminal nodes. These outcome variables were each selected to reflect different aspects of the algorithm's performance. The number of terminal nodes serves as an indicator of the ability of MBRP to correctly identify the number of distinct partitioning groups. When the level of partition DIF was simulated to be 0, we would expect the number of terminal nodes to be 1 as the groups are not distinct on the model parameters. Any difference would reflect Type I errors. On the other hand, when the level of partition DIF was greater than 0, we would expect MBRP to recover the same number of terminal nodes as there were partitioning groups (i.e., 2 terminal nodes in the 2 groups condition and 10 terminal nodes in the 10 groups condition). RMS was calculated for each of the model parameters in (3), and takes the form:

$$\sqrt{\frac{\sum_{i=1}^{N}\left(\beta_{jn}-\bar{\beta}_{j}\right)^{2}}{N}} \qquad (4)$$

where, $N =$ Number of terminal nodes; $\beta_{jn} =$ Estimate for parameter j in terminal node n; and $\bar{\beta}_j =$ Mean of parameter j estimates across the N terminal nodes.

RMS reflects the degree of difference across the terminal nodes in terms of the parameter estimate values. Larger RMS values correspond to greater differences in the values of a particular parameter estimate across the terminal nodes. Therefore, we would expect RMS to exhibit differences in value for the parameter manipulated to differ among the partitioning groups in this study, β_2, when the partition DIF was not 0. In addition, we would expect RMS for the other model parameters to be unchanged (except due to sampling variability) across other study conditions, reflecting the fact that no partitioning group differences were simulated for them.

All of the manipulated factors described above were completely crossed. Simulations were conducted using the mob function in the R software system (R Core Development Team, 2014), which is a part of the party package. In order to ascertain which of the manipulated factors, and their interactions, contributed to differences in the outcome variables, analysis of variance (ANOVA) was used, as suggested for such work (e.g., Boomsma, 2013; Paxton, Curran, Bollen, Kirby, & Chin, 2001). For a main or interaction effect to be considered important, it had to be both statistically significant and have an effect size (ω^2) value of 0.1 or greater. This latter condition was included to ensure that only those effects that contributed at least 10% to the variance of the study outcomes were discussed in detail.

Results

No Differences in the Levels of Partition Group DIF Present

ANOVA results indicated that two terms were statistically significantly related to the number of terminal nodes identified by the MBRP algorithm: the number of partition groups ($F_{1,94595} = 12.03, p = 0.0005, \omega^2 = 0.0001$) and the partition sample size by partition sample size ratio ($F_{2,94595} = 3.57, p = 0.0281, \omega^2 = 0.00008$). As can be seen from the effect size estimates, neither term accounted for even 1% of the variance in the number of nodes. The mean number of terminal nodes for 2 and 10 partition groups were 1.1 in each case. The distribution of the number of nodes by the number of partition groups appears in Figure 2. For both numbers of partition groups, more than 90% of the replications correctly did not split the data, reflecting the fact that no differences in partition group DIF levels were simulated in the population. For 2 partition groups, 2 terminal nodes was the maximum number found by the algorithm; whereas for 10 partition groups, the MBRP algorithm settled on 3 terminal nodes in 0.68% of the replications and 4 terminal nodes in 0.05% of replications. Table 2 includes the mean number of terminal nodes by partition sample size and partition sample size ratio. Under all combinations of conditions, the mean number of terminal nodes was approximately 1.1, which, again, would be expected given that the data were simulated to have no partition group differences.

Differences in the Level of Partition Group DIF Present

When the level of DIF was simulated to differ among the partition groups, ANOVA identified the 2-way interaction of difference in partition group DIF by number of partition groups ($F_{2,376622} = 75930.5.1, p < 0.0001, \omega^2 = 0.326$) to be the highest-order term that was statistically significant and accounted for at least 10% of the variance in the number of terminal nodes. In addition, the interaction of partition group sample size by number of partition groups was also statistically significant and accounted for at least 10% of the variance in the number of terminal nodes ($F_{2,376622} = 17651.1, p < 0.0001, \omega^2 = 0.101$).

The mean number of terminal nodes by the number of partition groups and the partition group DIF appears in Table 3. When there were 2 partition groups simulated to differ in the level of DIF, the mean number of terminal nodes increased concomitantly with the level of partition DIF. When partition DIF was simulated to be small (0.2), the mean number of terminal nodes was 1.42; whereas when it was simulated to be large (0.8), the mean number of terminal nodes was 1.98. In other words, for larger values of the difference in the level of partition DIF, the number of terminal nodes was closer to the actual simulated number of 2. A similar pattern was evident when there were 10 partitioning groups, with more terminal nodes identified by the MBRP algorithm for a greater degree of partition group DIF. The number of terminal nodes in the 10 groups case ranged from a low of 4.26 in the 0.2 partition DIF condition to 8.29 in the 0.8 condition. Another way in which to consider these results is in terms of the ratio of the mean number of terminal nodes to the actual number of partition groups. For example, when partition DIF was 0.2 in the 2 partition groups case, the ratio of the mean number of terminal nodes identified by the algorithm (1.42) to the actual number of group present (2) was 0.71. On the other hand, for 0.8, this ratio was 0.99. With respect to 10 partition groups, the ratio of extracted terminal nodes to actual nodes was 0.426 (4.26/10) in the partition DIF of 0.2 condition and 0.829 in the 0.8 case. Thus, the MBRP algorithm appears more accurate in terms of identifying the number of terminal nodes for 2 partitioning groups than for 10.

The distribution of the number of terminal nodes by the number of partition groups and the degree of partition group DIF appears in Table 4. When 2 partition groups were present and the difference in the level of DIF was 0.2, the MBRP algorithm correctly identified the presence of 2 groups in 41.8% of replications, and mistakenly found 1 node in the other 58.2% of cases. For greater levels of partition DIF, the correct result of 2 partitions occurred in 83.1% and 97.8% of the replications for 0.4 and 0.8, respectively. When 10 groups were simulated in the population and the level of partition DIF was 0.2, 49.3% of replications resulted in 4 terminal nodes being identified by MBRP, with another 27.5% identifying 5 terminal nodes. As was true for the 2 group case, when the separation among the partitioning groups increased in value, the number of recovered terminal nodes increased as well. For a difference in DIF of 0.4, 73.8% of the replications were identified as having between 5 and 8 different nodes; whereas for a difference in DIF of 0.8, 80.4% of replications yielded between 7 and 9 terminal nodes with an additional 14.4% having 10 terminal nodes.

In order to ascertain how splits were being made by the MBRP algorithm, the RMS taken across replications for each term in the model was examined. This statistic was calculated for each predictor variable for each replication across the terminal nodes. Larger values of RMS indicate a greater difference

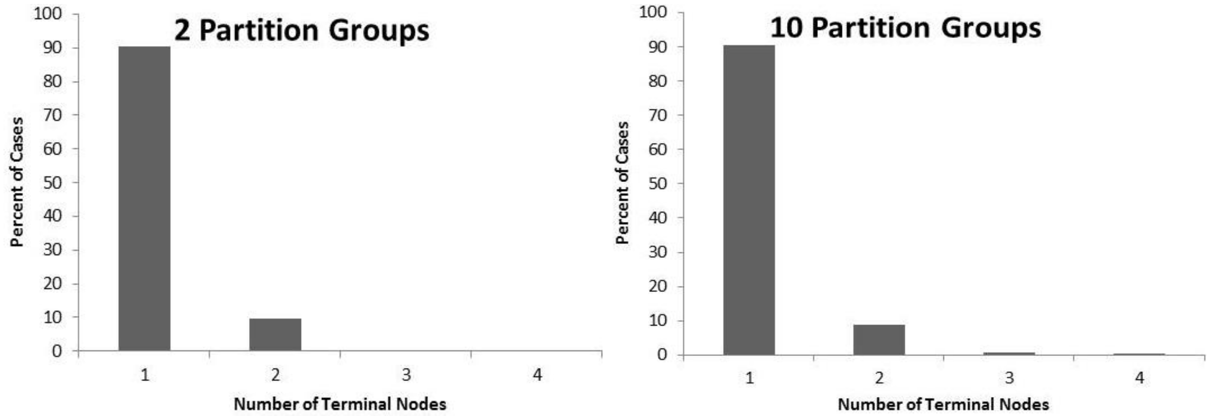

Figure 2. Percent of the number of terminal nodes recovered by MBRP by number of partition groups when no partition group DIF was present.

Table 2. Mean Number of Terminal Nodes by Partition Group Sample Size and Partition Group Sample Size Ratio in the No Partition Group Difference Condition

Partition N	Partition N Ratio	Mean Number of Terminal Nodes
250	Equal	1.10
250	Unequal	1.10
500	Equal	1.09
500	Unequal	1.10
1000	Equal	1.10
1000	Unequal	1.09

Table 3. Mean Number of Terminal Nodes by Number of Partition Groups and Partition Group DIF in the Partition Group Difference Condition

Number of Partitions	Partition DIF	Mean Number of Terminal Nodes
2	0.2	1.42
	0.4	1.83
	0.8	1.98
10	0.2	4.26
	0.4	6.57
	0.8	8.29

in the parameter estimate for the particular variable across the various terminal nodes in a solution. Therefore, this statistic can be helpful in explaining which parameters were contributing most to the splits yielding the terminal nodes. Greater differences in RMS across conditions indicate that the parameter estimates differed more across the terminal nodes for the specific condition. Thus, larger changes in the values of RMS across conditions, for the LR model parameters that are most strongly influenced by specific manipulated study conditions, are expected. Given that the focus of the study was on β_2 in model (3), we expect for it to exhibit the largest changes in RMS values across the simulated conditions.

With respect to the RMS of the group variable, which was manipulated to differ across the partitions, the 3-way interaction of the number of partition groups by the partition group ratio by the difference in partition group DIF was the highest-order statistically significant term in the ANOVA model $(F_{1,168} = 27.29, p < 0.0001, \omega^2 = 0.125)$. All other manipulated terms were either statistically significant, but subsumed under this 3-way interaction, or were not statistically significant while accounting for at least 10% of the variance in the outcome variable. The RMS by number of partition groups, partition group sample size ratio, and differences among the partition groups appear in Table 5. When the partition groups were of the same size and there were 2 groups in the population, RMS for $\beta_1$and β_3 in the LR model were not influenced by the level of partition DIF. In other words, increasing differences among the partition groups did not influence the estimation of either of these parameters. In contrast, increasing partition DIF was associated with an increase in RMS for the β_2 and β_0 estimates. This result was expected for β_2 as the increasing difference in the *cat* effect across partitions should result in greater differences in the

Table 4. Percent of Number of Terminal Nodes Recovered by the MBRP Algorithm, by Number of Partition Groups and Partition Group DIF in the Partition Group Difference Condition

Recovered Nodes	2 Partition Groups	10 Partition Groups
Difference Between Partition Group DIF = 0.2		
1	58.2	0
2	41.8	0.7
3	0	15.4
4	0	49.3
5	0	27.5
6	0	6.2
7	0	0.8
8	0	0.1
9	0	0
10	0	0
Difference Between Partition Group DIF = 0.4		
1	16.9	0.7
2	83.1	0.3
3	0	1.8
4	0	3.6
5	0	14.1
6	0	29.8
7	0	28.9
8	0	10.0
9	0	4.1
10	0	6.8
Difference Between Partition Group DIF = 0.8		
1	2.2	0
2	97.8	0
3	0	0
4	0	0
5	0	0.4
6	0	5.0
7	0	18.9
8	0	30.9
9	0	30.6
10	0	14.4

Table 5. RMS for Coefficient Estimates of the Intercept, Cont, Cat, and Interaction Terms by Partition Group Size Ratio, Number of Partition Groups, and Partition Group DIF in the Partition Group Difference Condition

Number of Partition Groups	Partition DIF	Intercept	Cont	Cat	Interaction
		Partition Sample Size Ratio Equal			
2	0.2	0.17	0.01	0.01	0.01
	0.4	0.20	0.01	0.17	0.01
	0.8	0.38	0.01	0.38	0.01
10	0.2	0.68	0.07	0.41	0.09
	0.4	1.04	0.06	0.73	0.08
	0.8	3.03	0.15	2.62	0.17
		Partition Sample Size Ratio Unequal			
2	0.2	0.20	0.01	0.07	0.01
	0.4	0.23	0.01	0.07	0.01
	0.8	0.39	0.01	0.11	0.01
10	0.2	0.75	0.08	0.77	0.10
	0.4	1.91	0.22	1.73	0.24
	0.8	3.48	0.41	3.79	0.33

estimates of this coefficient across partitions, which would, in turn, be reflected in a larger RMS value. On the other hand, increasing partition DIF for the *cat* variable was not expected to impact the estimates of β_0. Results in Table 5 also show that for 2 partition groups that were of unequal size, the impact of partition DIF on RMS was much more muted than in the equal condition. The results for the β_0, β_1, and β_3 estimates were very similar in both the equal and unequal conditions. In the 10 partition equal groups size case, RMS for β_0 and β_2 both increased concomitantly with increases in the difference of the group variable across partitions. In addition, RMS for β_1 and β_3 were both larger in the 0.8 partition DIF condition than in the 0.2 and 0.4 conditions. When there were 10 partition groups of unequal size, RMS for all of the model coefficients was greater compared to the equal group size situation, and RMS increased along with increases in partition group DIF. This result represented a different pattern for β_1 and β_3 RMS values than was seen in the equal group size condition.

Discussion

The goal of this study was to investigate the performance of the MRBP algorithm as a data analysis tool in the presence of multiple organizing units of individuals such as nations or states. This method can have application in a wide variety of contexts, including DIF assessment, which was the focus of this work. However, MBRP is clearly not limited to the measurement and psychometrics context and can be easily employed in a range of research scenarios. Based on the results presented above, MBRP was able to accurately identify the case where the partitioning groups did not differ on any LR model parameters approximately 90% of the time. In other words, whether there were 2 or 10 separate groups or when the groups were simulated to have the same model parameters, MBRP correctly identified a single terminal node in 90% of the cases. When the 2 partition groups' model parameters differed by 0.8, MBRP correctly identified 2 terminal nodes nearly 100% of the time. In contrast, when these model parameters differed by 0.4, the algorithm only separated the partition groups from one another in approximately 42% of cases. For 10 partition groups, MBRP rarely recovered the correct number of terminal nodes. However, as the group separation on the target parameter increased in value, the number of recovered terminal nodes increased concomitantly by becoming closer to the correct number of 10. For the greatest degree of group separation; however, the correct number of 10 terminal nodes was recovered only 14.4% of the time, indicating less accuracy with more model complexity.

While on the one hand, these results may not appear overly promising for the MBRP method, particularly when there are many partition groups, it should be noted that the models only differed on a single model parameter value. In other words, the algorithm was attempting to differentiate as many as 10 groups based on a model that differed for only a single parameter. In this light, the performance of MBRP might be viewed somewhat more positively. Put another way, even when only a single model parameter differed among the partitioning groups by as little as 0.4, MBRP was able to correctly separate 2 groups

83% of the time, and was able to identify the presence of 6 or more groups in approximately 80% of cases when 10 were in fact present. When the single coefficient differed by 0.8, the algorithm correctly identified the presence of 2 groups in nearly every case and found 6 or more groups 91% of the time. Clearly, further simulation research should be conducted with this methodology in which more model parameters are allowed to vary among the partitioning groups. Furthermore, applied examples following these more complex models can be examined, where multiple variables are investigated to determine how more complex models can assist in understanding performance across many groups at once.

Another finding of note in the current study was that partition group differences in the categorical predictor coefficient was associated with differences in the intercept estimate across the terminal nodes, as reflected by RMS. For all model terms, except for the categorical predictor, the model parameter estimates in the terminal nodes should have been very similar to one another, regardless of the level of partition DIF. However, as was noted in the results, this was not found to be true in all situations, particularly for 10 partition groups. In addition, when the groups were of unequal size, this perturbation in the intercept estimate was exacerbated (i.e., the RMS for the intercept increased in value as partition DIF on the coefficient for the categorical variable increased) and also became apparent for the other model parameters in the 10 groups case. This finding that MBRP had difficulty in accurately reflecting a lack of partition group difference in model parameter estimates for a large number of unequally sized groups is in keeping with prior research, indicating that CART, in particular, exhibits some problems with unequally sized partitioning groups (Holden, Finch, & Kelley, 2011).

Complex sampling and the need to maximize information obtained from statistical models, while properly accounting for data structure, will only gain in importance as complex databases are used more frequently. In the context of this study, DIF detection will also remain essential in test development as scales are constructed to facilitate comparisons across organizing units such as nations. Numerous statistical methods have been developed, yet few are tested in the presence of complex sampling and with multiple variables. This issue is of particular import given the increasing interest in large multi-state and multinational databases.

References

Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling, 20*, 518-540.

Brieman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* New York: Wadsworth.

Holden, J. E., Finch, W. H., & Kelley, K. (2011). A comparison of two-group classification methods. *Educational and Psychological Measurement, 71*, 663-683.

Kim, H., & Loh, W. Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association, 96*, 589-604.

Organization for Economic Co-operation and Development. (2014). *TALIS 2013 results: An international perspective on teaching and learning*. Retrieved from http://dx.doi.org/10.1787/9789264196261-en

Organization for Economic Co-operation and Development. (2013). *PISA 2012 results: What students know and can do – Student performance in mathematics, reading and science (Volume I)*. Retrieved from http://dx.doi.org/10.1787/9789264201118-en

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8*, 287-312.

R Core Development Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: Author.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Zeileis. A., Hothorn. T., Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics, 17*, 492–514.

Send correspondence to: W. Holmes Finch
Ball State University
Email: whfinch@bsu.edu