International Journal of Testing, 16: 21–53, 2016
Copyright © Taylor & Francis Group, LLC
ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305058.2015.1039644

Routledge
Taylor & Francis Group

Recursive Partitioning to Identify Potential Causes of Differential Item Functioning in Cross-National Data

W. Holmes Finch and Maria E. Hernández Finch
Department of Educational Psychology, Ball State University

Brian F. French
Department of Educational Leadership and Counseling Psychology, Washington State University

Differential item functioning (DIF) assessment is key in score validation. When DIF is present scores may not accurately reflect the construct of interest for some groups of examinees, leading to incorrect conclusions from the scores. Given rising immigration, and the increased reliance of educational policymakers on cross-national assessments such as Programme for International Student Assessment, Trends in International Mathematics and Science Study, and Progress in International Reading Literacy Study (PIRLS), DIF with regard to native language is of particular interest in this context. However, given differences in language and cultures, assuming similar cross-national DIF may lead to mistaken assumptions about the impact of immigration status, and native language on test performance. The purpose of this study was to use model-based recursive partitioning (MBRP) to investigate uniform DIF in PIRLS items across European nations. Results demonstrated that DIF based on mother's language was present for several items on a PIRLS assessment, but that the patterns of DIF were not the same across all nations.

Keywords: *classification and regression trees, differential item functioning, model based recursive partitioning, PIRLS*

Differential item functioning (DIF) detection is central to validity arguments to support inferences from assessment results (American Education Research Association [AERA], American Psychological Association, & National Council on

Correspondence should be sent to W. Holmes Finch, Department of Educational Psychology, Ball State University, 521 Teachers College, Muncie, IN 47306, USA. E-mail: whfinch@bsu.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hijt.

Measurement in Education, 2014). DIF represents a lack of construct comparability among examinees belonging to different subgroups in the population, such as males and females, or native speakers of the testing language, and nonnative speakers. Note that the terms "assessment" and "test" are used interchangeably in this article. In other words, DIF occurs when members of two distinct groups have different probabilities of endorsing an item, when they are matched on the overall trait being measured by the assessment. For example, gender DIF occurs for an item on a reading assessment if, after matching boys and girls on their reading aptitude score, the probabilities of a correct response differed for boys and girls. Such a difference in item level performance is problematic as it indicates that the assessment is not performing comparably for the two groups of individuals, implying that the construct is not measured similarly for boys and girls (Camilli & Shepard, 1994).

DIF detection is typically conducted in the context of a single population where examinees can be differentiated by an easily measureable variable such as gender or ethnicity. In some instances, such easily defined variables are not available, and a mixture model approach is used to ascertain whether DIF is present in terms of some heretofore unobserved grouping in the population (Cohen & Bolt, 2005). In either case, the sample as a whole is typically drawn from what is viewed to be a single population, such as students in the US education system. However, in the context of international testing, the sample might be taken from a number of different nations across the world. Examples include the Programme for International Student Assessment (PISA; Organisation for Economic Co-Operation and Development [OECD], 2013), Trends in International Mathematics and Science Study (TIMMS; International Association for the Evaluation of Educational Achievement [IEA], 2011), and Progress in International Reading Literacy Study (PIRLS; IEA, 2011), among others. In the context of these large cross-national studies, examinees are sampled from several nations and are given a common set of assessments. Because of the diversity of cultures, educational systems, and experiences of examinees, DIF detection in this context may be a great deal more complex compared to what is typically described in the literature. For example, a researcher interested in gender-based DIF on reading items in the PIRLS initiative would need to consider the very different cultural and educational environments of diverse nations (e.g., United States contrasted with Saudi Arabia). For more than a very small number of nations such an analysis could be extremely difficult.

In addition to problems associated with complex data sets involving individuals sampled from multiple nations (e.g. PISA, PIRLS, TIMMS), or states (e.g. National Assessment of Educational Progress [NAEP]), there has also been increased interest in the DIF literature on moving beyond mere identification of DIF to its explanation (Albano & Rodriguez, 2013). The goal in this case would be to use statistical techniques, such as multilevel models, in conjunction with a content review by experts for items identified as exhibiting DIF, in order to gain insights into the reasons that DIF might be present. Typically, the statistical analysis aspect of DIF analysis is completed using models that include the variable of interest with respect to DIF (e.g., examinee gender), as well as additional variables that are hypothesized to be associated with the occurrence of DIF with respect to the target variable. As an example, prior research has studied the extent to which opportunity to learn is associated with DIF in terms of gender (Albano & Rodriguez, 2013; Cheong, 2006; Clauser, Nungester, & Swaminathan, 1996). Other studies have also attempted to move beyond simply identifying the presence of DIF, but also providing explanations as to its causes as well, using a combination of statistical analysis and substantive review of the items identified statistically as exhibiting DIF (e.g., Balluerka, Gorostiaga, Gómez-Benito, & Hidalgo, 2010; Walker & Beretvas, 2006). In addition to methods such as logistic regression, the Mantel-Haenszel test, and SIBTEST, prior work in this area has also employed mixture item response theory (IRT) models to investigate patterns of DIF for international assessments, such as PIRLS (Oliveri, Ercikan, & Zumbo, 2013; de Ayala, Kim, Stapleton, & Dayton, 2003), and multilevel latent class models have been applied to clustered data (Vermunt, 2008).

Given potential problems associated with DIF analyses across a large number of organizing entities such as nations, as well as the increasing recognition as to the importance of providing explanations for DIF, the current study seeks to demonstrate the use of model based recursive partitioning (MBRP) as a potentially useful tool for researchers and measurement specialists interested in assessing and understanding DIF. MBRP makes it possible for DIF detection to be conducted using a well-known and effective tool, logistic regression, while allowing DIF model coefficients to differ across these organizing entities, where appropriate. In addition, multiple variables of interest can be included in the logistic regression model, thus providing deeper insights into correlates of DIF, and thereby giving experts additional information regarding potential aspects of the items that should be accorded special attention in substantive reviews used to investigate the causes of DIF. In addition, the MBRP framework allows for more than one partitioning variable (to be explained next), which can provide the researcher with additional information regarding instances of DIF, and how it manifests itself differently in different organizing entities. In short, the flexibility of the MBRP approach allows for the inclusion of a wide variety of variables in the model, which in turn can provide experts reviewing DIF items with greater information regarding potential causes of DIF, which they can in turn bring to bear in their review of the item content vis-à-vis the target variable(s) for which it was found to be present statistically.

The remainder of this manuscript is organized as follows. First, we describe the specific problem of interest here, namely the role that examinees' native language status has on item performance, when controlling for the amount of educational resources that are available in the home. Next, we describe classification and regression trees, upon which MBRP is based. Discussion of MBRP is then followed by an explanation of the study's goals, the methods used to address those goals, and the results of the analyses. Finally, we discuss the results in order to highlight both the substantive findings regarding DIF and performance on reading items, and how they might differ across a group of nations, as well as implications of this study for practice, particularly with reference to the benefits and drawbacks of using the MBRP approach.

Native Language Status and Reading Test Performance

Language status is a confounding variable and a source of measurement error on standardized tests (Abedi, 2002; AERA et al., 2014). The primary concern is whether the assessment items are measuring only the intended content, or in part, are also measuring language proficiency. In research conducted in the United States, English learners (ELs) showed greater achievement gaps on such measures when the task's language demands are higher (such as in reading assessments) and in the upper grade levels when item complexity increases (Abedi, 2002). Koo, Becker, and Kim (2014) conducted a meta-analysis of DIF to explore four types of reading comprehension test questions on Florida's high-stakes assessment (1. phrases-in-context, 2. main idea, 3. cause-effect, and 4. evaluation) with three groups of EL students (Caucasian EL, Hispanic/Latino EL, and Asian EL) and a non-EL referent group. In the third grade, when controlling for overall reading ability, phrases-in-context vocabulary questions showed DIF for EL students. When controlling for reading ability with tenth graders, ELs generally outperformed the non-EL group on evaluation items. At that grade level, DIF was detected for Caucasian ELs on main idea questions, with results favoring the non-EL group.

Socioeconomic status (SES; Abedi 2002; Abedi & Lord, 2001) and parent education levels affect performance on high-stakes tests, with EL students being disproportionally represented in lower levels of both factors (Abedi, 2002). Recent longitudinal research (Herbers et al., 2012) suggested that early reading success is comparatively more important for those students who are at risk by virtue of high residential mobility and/or lower SES in predicting later achievement and learning. In their review, Heath, Rothon, and Kipli (2008) asserted that parental knowledge of the language of the new country and even parental knowledge of the educational system was related to performance on standardized tests in various European countries. It is important to note that recent research supports that some forms of bilingualism are an academic and social boon (see Han, 2012).

Taken together, these prior research results informed the direction of the current study, which investigated DIF with respect to maternal language status across several nations. In addition to examining the presence of DIF for reading items cross-nationally, the impact of SES factors such as parental education and occupation status, among others, were controlled for using the composite variable available educational resources, which is described next. This focus allowed the study to be aligned with recent attention on cross-national DIF assessment (Glas & Jehangir, 2014).

Classification and Regression Trees

An analytic approach gaining attention in the general statistics literature, and which may be a useful tool for explanatory DIF analysis, particularly in the presence of multiple organizational units of interest (e.g., nations, states, school districts), is MBRP (Su, Wang, & Fan, 2004), which is based on the classification and regression tree (CART) methodology (Brieman, Friedman, Stone, & Olshen, 1984). In the case of CART, a researcher specifies a prediction model in which one variable serves as the outcome, and a set of variables serve as predictors of this outcome. To be clear, consider the case where the outcome is categorical in nature (e.g., dichotomously scored items), and the predictors are a mixture of categorical (e.g., gender) and continuous (e.g., family SES placed on a standard normal scale) variables.

CART begins by placing all members of the sample in a single grouping called node 1, and searches the entire sample for the binary partition among the predictors that results in the most homogeneous split possible, in terms of the outcome variable. This split creates daughter nodes, and individuals are placed into the appropriate such node based on his or her predictor variable value. For example, if the optimal split is on gender, then males are placed into one daughter node (e.g., node 2), and females are placed into the other daughter node (e.g., node 3). The CART algorithm will next investigate potential splits in each of the daughter nodes, again with the goal of creating two even more homogeneous nodes with respect to the outcome variable, in this case response to the item. Continuing with the example, assume that the optimal split for node 2 is on SES at a value of -1 such that individuals with a family SES of less than -1 are placed in daughter node 4 and those with a SES of -1 or higher are placed in node 5. Returning to node 3 (females), assume that the optimal split is on age at a value of 7 years, where those younger than 7 years of age are in node 6 and those older than 7 are placed in node 7. This partitioning will continue until further separation does not yield increased homogeneity in the dependent variable, resulting in a decision tree that differentiates the groups. We will now illustrate this process graphically for MBRP.

Model-Based Recursive Partitioning

MBRP uses the same partitioning methodology as CART. Indeed, CART can be seen as a special case of MBRP where the model of interest is simply the mean of the dependent variable (Chan & Loh, 2004; Gama, 2004; Kim & Loh, 2001; Loh, 2002; Zeileis, Hothorn, & Hornik, 2008). However, whereas CART builds a tree so as to maximize partitioning group differences on the mean of the dependent variable, MBRP maximizes partitioning group differences on a model such as logistic regression. As described in Zeileis and colleagues, this recursive partitioning approach takes the following steps:

1. Fit the model of choice (e.g., logistic regression) to all observations in the current node (e.g., node 1).
2. Assess parameter stability for each of the independent variables in the model, and select the independent variable with the highest such instability. If stability is present, the algorithm stops. If stability is not present, proceed to step 3.
3. Compute the split point of the model parameter values that optimizes partitioning group separation.
4. Split the node based on this split point to create two daughter nodes.
5. Repeat steps 1 through 4 until parameter stability is achieved.

Step 1 of the algorithm simply involves fitting the model of interest in the normal fashion. In the case of uniform DIF assessment, this model would be logistic regression (LR) with the item response as the dependent variable, and the independent variables being the score on the test and the variable for which we wish to assess DIF, such as gender (Swaminathan & Rogers, 1990). This model is expressed as

$$ln\left(\frac{\pi\,(x)}{1-\pi\,(x)}\right)=\beta_0+\beta_1 score+\beta_2 gender. \qquad (1)$$

Here, $\pi\,(x)$ is the probability of a correct item response, β_0 is the model intercept, β_1 is the coefficient for total test score, and β_2 is the coefficient for gender, which assesses whether uniform DIF based on gender is present for the item. In the context of MBRP, the goal of step 2 is to determine whether the model parameter estimates obtained in step 1 are stable across all combinations of the partitioning variable of interest. In the current example, let us assume that the partitioning variable of interest is nation, and that item response data have been collected from examinees representing 20 different nations. In step 2, the MBRP algorithm would assess whether the coefficients for the independent variables test score and gender in (1) are stable across all possible partitions of the 20 nations. The stability assessment is made using the test statistic $\lambda_{\chi^2}\left(W_j\right)$, which is asymptotically distributed as χ^2 with $k*(C-1)$ degrees of freedom, where k is the number of model coefficients, and C is the number of categories in the partitioning variable (e.g., nations). In the case of (1), k would be 3. This statistic takes the form:

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

In Equation (2), j refers to the partitioning variable used to make the split. If only one partitioning variable is used (e.g., nation is the only partitioning variable), then the statistic in Equation (2) will only be calculated for $j = 1$. The term $W_j\left(\frac{i}{n}\right)$ is defined as:

$$W_j\left(\frac{z_j}{n}\right) = \hat{S}^{-1/2} n^{-1/2} \sum_{j=1}^{z_j} \psi_{z_j}. \tag{3}$$

The value ψ_{z_j} is the score function estimated for the model in Equation (1) for the partition z_j. Here, the z_j partitions simply refer to the grouping of the observations by the partitioning variable (e.g., nation in this study). In other words, for each partition there exists a score statistic z_j, and these are summed across all of the partitions. This sum is then scaled by the total sample size, n, and the covariance matrix of the score functions, S. The value I_c is the number of individuals in category c of the partitioning variable and I is the total number of individuals in the node. The value $\Delta_{I_c} W_j$ measures whether there are systematic fluctuations in the score function associated with the model of interest, LR in this case, across the C levels of the partitioning variable (e.g., nations); that is, it is the sum of $W_j\left(\frac{z_j}{n}\right)$ across C. If the model parameters are comparable for all levels of C, the score function should fluctuate randomly around its mean of 0 (Zeileis et al., 2008). Thus, systematic fluctuations in the score function from one level of C to another would indicate instability in the model parameters from Equation (1), such that different model forms are needed for different combinations of the partitioning variable. Calculation of the statistic in Equation (2) for a given node requires that the model in Equation (1) be fit once for all individuals in the node, after which the score function values are simply reordered and aggregated to calculate $\lambda_{\chi^2}\left(W_j\right)$ for each possible combination of the C levels of the partitioning variable.

A statistically significant $\lambda_{\chi^2}\left(W_j\right)$ value for a node indicates that the model parameters are not equal across the levels of C in the node (i.e., are not stable). In that case, a binary split is made based on the partitioning variable. In order to determine where the split will be made, an exhaustive search of all possible combinations of the partitioning variable (the Nation variable in this study) is made, where for each partition the model in Equation (1) is fit. For each partition, the sum of the score function is calculated across individuals within each node, and the split that minimizes the score function sum across the two daughter nodes is selected. As noted, these steps are repeated for each of the resultant daughter nodes. Partitioning stops when parameter stability has been reached. That is, the recursive partitioning algorithm stops when the $\lambda_{\chi^2}\left(W_j\right)$ is not statistically significant, indicating that within the nodes the model parameters are consistent across levels of the partitioning variable.

An important issue when working with recursive partitioning models such as the MBRP one used here is the problem of the model overfitting the data (Hothorn, Hornik, & Zeileis, 2006). Overfitting simply means that the model parameters are so closely tied to the sample upon which the model was built that they are not generalizable to another sample from the same population. In order to avoid overfitting with MBRP, it is recommended that the researcher first grow a full size tree based on the algorithm described previously, and ascertain the value of the Akaike Information Criterion (AIC). AIC is a measure of relative model fit that incorporates unexplained variance in the data and applies a model complexity parameter (Akaike, 1974). A model with a smaller value of AIC is considered preferable to a model with a larger AIC. The full sized tree should be inspected for overfitting using an exploratory technique known as pruning. With pruning, the researcher removes a single terminal node and calculates the AIC for this reduced model. If the AIC is smaller than that for the full tree, the terminal node should be left off of the tree, whereas if the AIC of the pruned tree is larger than the value for the full tree, then the terminal node should be kept in the tree. Each terminal node should be assessed in this manner in order to determine which should remain in the tree and which should be removed. The order in which the terminal nodes should be assessed is based on the *p*-values associated with the stability test statistics described previously. The split associated with the largest *p*-value would be the first to be checked in the pruning process, followed by the split with the next largest *p*-value, and so on until all terminal nodes have been assessed (Zeileis & Hornik, 2007).

There has not been a great deal of research conducted examining the power and Type I error rate of the $\lambda_{\chi^2}\left(W_j\right)$ test in the context of MBRP. One simulation study that has examined these properties was conducted by Zeileis and colleagues (2008). These authors simulated data containing a simple linear regression equation, and standard normal variables for partitioning the sample, with a total sample size of 500. They found that when the data were simulated under the null hypothesis of no partitioning in the population, the Type I error rate was 0.038. Conversely, when there was partitioning simulated in the population, the power of MBRP was 1.000 across simulated conditions for detecting change across the partitioning variable in either the model intercept or the slope. Finally, Zeileis and colleagues found that MBRP split the sample more frequently than it should at a rate of 0.067. In other words, when 1 split was simulated, MBRP identified 2 or more splits in approximately 6.7% of cases. While no other study appears to have directly investigated the power and Type I error rates of the test statistic with respect to MBRP, Potts and Sammut (2005) did examine the performance of $\lambda_{\chi^2}\left(W_j\right)$ in the context of CART and found that it maintained a 0.05 Type I error rate when no partitioning was simulated in the population. Taken together, the results of these studies would appear to support the Type I error and power rates of $\lambda_{\chi^2}\left(W_j\right)$ in the context of MBRP, although clearly more work in this regard is needed.

Goals of the Current Study

For researchers interested in cross-national studies involving large-scale assessments, a primary question of interest revolves around the extent to which examinee performance, and the factors that impact it, are similar and different from one country to the next. = Quite often, such studies focus on performance at the total test score level (e.g., OECD, 2010a, 2010b). However, given that total test scores are based on performance on individual items, it may also be of interest to determine the extent to which such items perform comparably for specific subgroups in the population. In particular, this study focuses on the issue of language spoken in the home, and whether it is associated with DIF for items on assessments used in the international reading assessment, PIRLS. Of particular interest in the current study is the extent to which a mother's primary language is associated with DIF on reading items from a PIRLS test, when controlling for the amount of educational resources available in the home, and whether these relationships are consistent across several nations. As noted, prior research within nations (e.g., the United States) has demonstrated that the language spoken in the home is associated with DIF, and that home resources are associated with performance on achievement tests. We wish to extend this within-nation work in order to ascertain whether such patterns are consistent cross-nationally, when controlling for resources in the home. This comparison allows for demonstrating MBRP within an applied and important context, taking advantage of existing large-scale data where results can have very real and practical applications for entire nations. Indeed, such results from these sources are used to directly inform educational policy.

To assess DIF and determine whether its patterns were consistent across nations, MBRP based on the widely used and popular LR model for DIF detection was used (Swaminathan & Rogers, 1990). With MBRP, we anticipated being able to identify whether DIF based on home language was present when controlling for available educational resources, and whether the nature of DIF takes a consistent form across nations. We could find no evidence of this method being used in this manner and believe this is its first application to the problem of DIF detection to assess whether the nature and presence of DIF is consistent across several organizing entities such as nations.

METHOD

Data for this study were drawn from the PIRLS research initiative, using the 2011 international database (IEA, 2011). For the purposes of this study, the 13 items

TABLE 1
Process of Comprehension, Item Discrimination, and Item Difficulty Values for the Items Comprising the Test

		Item Parameter Estimates	
Item	Process of Comprehension	Discrimination	Difficulty
1	Interpret and integrate ideas and information	1.58	−1.58
2	Make straightforward inferences	0.85	−2.20
3	Make straightforward inferences	0.97	−1.73
4	Focus on and retrieve explicitly stated information	1.50	−2.06
5	Make straightforward inferences	1.09	−1.88
6	Make straightforward inferences	1.05	−0.86
7	Interpret and integrate ideas and information	0.96	0.45
8	Make straightforward inferences	1.50	−0.88
9	Focus on and retrieve explicitly stated information	1.79	−0.96
10	Make straightforward inferences	1.19	−2.22
11	Examine and evaluate content, language, and textual elements	0.64	−0.97
12	Interpret and integrate ideas and information	1.25	0.96
13	Examine and evaluate content, language, and textual elements	1.02	−1.24

associated with the Flowers assessment were assessed for uniform DIF. The Flowers assessment consists of a brief passage describing a flower garden being grown on the roof of a building, and then asks the examinees to respond to 13 multiple-choice items designed to measure various aspects of reading comprehension for the passage. The items were dichotomously scored as correct (1) or incorrect (0). The comprehension processes that were measured by each item appear in Table 1, along with the item discrimination and difficulty parameter estimates for the entire studied sample based on the 2-parameter logistic (2PL) model. Both the 2PL and

TABLE 2
Percent of Households in Which Mother Spoke Language of Test, child's language prior to School Is Language of the Test, and the Language of the Test is Always Spoken in the Home, by Nation

Nation (N)	Mother's Language Is Language of Test	Child's Language Prior to School Is Language of Test	Language of Test Is Always Spoken in the Home
Ireland (401)	90.5%	92.6%	82.8%
Italy (506)	91.8%	94.1%	79.1%
Lithuania (700)	96.5%	98.5%	83.6%
Netherlands (395)	95.7%	96.8%	78.5%
Norway (465)	91.7%	96.3%	80.2%
Poland (818)	99.6%	99.3%	90.2%
Portugal (597)	97.8%	98.1%	89.7%
Romania (715)	96.4%	97.3%	90.9%
Russian Federation (713)	94.5%	96.2%	85.2%
Slovak Republic (933)	94.3%	97.2%	79.8%
Sweden (555)	85.3%	93.0%	76.1%
Total (6798)	94.2%	96.4%	83.4%

3-parameter logistic (3PL) models were fit to the Flowers data in order to obtain item parameter estimates. The determination as to the optimal model was made by comparing values of the AIC for each, as well as through the likelihood ratio test (de Ayala, 2009). The AIC for the 3PL was 110091.2, whereas AIC for the 2PL was 100873.6, indicating that the 2PL was preferable once model complexity was taken into account. In addition, the likelihood ratio test comparing the fit of the two models was not statistically significant ($\alpha = 0.05$), meaning that the more complex 3PL did not yield superior fit to the data despite its inclusion of pseudo-chance parameters, thereby providing further evidence as to the optimality of the simpler 2PL model.

The study sample consisted of 6798 examinees from the 11 nations, displayed in Table 2. We restricted the analysis to these 11 nations for three reasons. First, we wanted to have a sample of individuals that while diverse in terms of language and relative wealth (see Herbers et al., 2012 for a discussion of the importance of SES on achievement), was homogenous enough to not confound comparisons to a great extent by other variables (e.g., individualistic vs. collectivist cultures). Thus, we felt focusing on European nations would provide that balance of heterogeneity. Second, Europe has recently experienced a bourgeoning upsurge in immigration (Heath et al., 2008), making it particularly interesting with respect to the substantive issue of home language. Our third reason for selecting these 11 nations was practical. All examinees were given the Flowers assessment. PIRLS uses a number of test booklets as a part of the assessment effort, and not all nations have examinees completing each of these booklets. Thus, results would not be confounded with differences across booklets. We were particularly interested in assessing DIF for the Flowers test because it was used in previous iterations of PIRLS, and had well-documented psychometric qualities. Together this allowed for comparisons of DIF results over time and can inform future PIRLS work in addition to the methodological information. In order to ensure that items within the Flowers assessment did not violate the IRT assumption of local independence, Yen's Q3 statistic was used (Yen, 1984). The commonly recommended cut value of 0.2 was employed, such that Q3 values of 0.2 or larger would indicate that an item pair was locally dependent (Yen, 1993). In this case, none of the item pairs had Q3 values above the cut value, with the largest being 0.156. Therefore, it was determined that the local independence assumption was satisfied.

LR was used to detect uniform DIF in terms of two variables: the language spoken in the home by the student's mother was the same as the language of the test, and the educational resources available in the home. The first variable was coded as either the mother spoke the language of the assessment in the home (1), or not (0). The second variable was a composite created using a partial credit IRT model based on five items from the PIRLS student and home questionnaires, including the number of books in the home (as reported by students and parents), the number of home study supports (e.g., number of books in the home, number of children's books in the home, internet connection, quiet study space), the highest level of education for either parent, and the highest level of occupation by either parent (Martin, Mullis, Foy, & Arora, 2012). Higher scale scores indicate the presence of more educational resources in the home.

Only uniform DIF was assessed for purposes of illustrating the MBRP method. LR can be used quite successfully to test for nonuniform DIF as well (Narayanan & Swaminathan, 1996). However, doing so in this context would lead to the estimation of a larger set of DIF estimates (uniform and nonuniform for each item), which in turn could lead to much more complex trees. The decision was made that such complexity might serve to make the results more difficult to understand, which would not serve the purpose of introducing the MBRP methodology for use in DIF assessment with clustered data. It is important to keep in mind, however, that LR is an effective tool for investigating nonuniform DIF, and that such investigations could be very important in contexts where MBRP would be useful. Thus, researchers interested in nonuniform DIF assessment with clustered data should keep in mind the potential for such analysis with MBRP.

For each of the 13 items, a LR model was fit to the data in the context of MBRP in order to determine (a) whether uniform DIF was present with respect to either mother's home language or available educational resources in the home, controlling for total score on the test (and controlling for one another as well), and (b) whether DIF presented itself differently across the nations of Europe. The LR model used as a part of the MBRP took the following form.

$$ln\left(\frac{\pi\left(x\right)}{1-\pi\left(x\right)}\right) = \beta_0 + \beta_1 score + \beta_2 Mother'slanguage$$
$$+\beta_3 Educationalresources. \quad (4)$$

Pruning using AIC to compare reduced tree models to the full model was conducted, as recommended in the literature and described previously. The MBRP models were fit using the mob function that is a part of the party library in the R software language (R Foundation for Statistical Computing, 2013). The mob function is an R function written specifically to carry out the type of parametric model–based recursive partitioning that is used in this study. It is a part of a larger suite of functions that are contained in the party library. Each of the functions in this library is designed to fit some type of recursive partitioning model, including both parametric and nonparametric, the latter of which includes classification and regression trees and random forests. For more information on the party library and its functions the interested reader is referred to Zeileis, Hothorn, and Hornik (2014).

An alternative approach for DIF testing when using data drawn from multiple nations is in the form of a multilevel LR model. This model has been shown to be an effective tool for DIF assessment when examinees are clustered, as is the case here (Cheong, 2006; French & Finch, 2010). In this context, the fact that item response data were obtained using individuals from different nations is accounted for with a two-level model, in which the first level represents the individual examinees, and the second level represents the nation in which each examinee resides. This multilevel LR model takes the form:

$$\text{Level 1}: y_{ij} = \beta_{0j} + \beta_{1j}x + \varepsilon_{ij} \quad (5)$$

$$\text{Level 2}: \beta_{0j} = \gamma_{00} + U_{0j} \quad (6)$$

$$\beta_{1j} = \gamma_{10} + \mathrm{U}_{1j} \quad (7)$$

Equation (4) takes a very similar form to Equation (3), with the addition that the subscripts *ij* refer to the *i*th individual in the *j*th cluster (nation). In addition, γ_{00} is the fixed portion of the intercept that remains constant across all clusters, and U_{0j} is the random portion of the intercept that varies from cluster to cluster. Finally, g_{ij} is the fixed portion of the slope relating the independent variable (e.g., mother's language status) to the dependent, and U_{1j} is the random portion of the slope, which varies among the clusters. This inclusion of the random coefficients term allows for different relationships between mother's language and the item response, while conditioning on total reading ability, and educational resources in the home. A statistically significant test for the fixed coefficient associated with mother's language would indicate the presence of uniform DIF across nations, while a statistically significant random coefficient term would mean that the nature of this DIF was different across nations. Such a result for either the intercept or the slope would not provide information about the nature of the relationships within specific nations, but would indicate whether such specific relationships differed across nations. In short, the model will not yield individual coefficients by nation or for groups of nations, but will simply indicate whether the model parameters differ from one nation to another.

RESULTS

Descriptive Statistics

Table 1 contains the item difficulty and discrimination parameter estimates based on a 2-PL IRT model for the 13 items that comprise the assessment, as well as the process of reading comprehension targeted by each item. An examination of these parameters reveals most of the items were relatively easy, with difficulty values less than 0 on an IRT scale (Camilli & Shepard, 1994). The exceptions to this result were items 7 and 12, both of which required examinees to interpret and integrate ideas and information. Of the 13 items, 9 had discrimination parameter estimates greater than 1, with only item 11 having a value less than 0.8. Table 2 includes the percent of households in which the mother spoke the language of the test, the child's language prior to attending school was the language of the test, and the percent of children for whom the language of the test was always spoken in the home. These descriptive statistics are reported by nation. Nations with the highest rates of mothers speaking the testing language in the home were Poland and Portugal, whereas those with the lowest such rates were Sweden, Ireland, Norway, and Italy. Similar patterns were in evidence for the percent of cases where the child spoke the language of the test prior to attending school, and whether the language of the test was always spoken in the home. For both of these variables, Poland and Portugal had among the highest rates, along with Lithuania (child speaks testing language prior to school) and Romania (language of test is always spoken in the home). The lowest percentages for the child's language prior to school being the testing language were found in Sweden and Ireland, and the lowest values for testing language always spoken in the home appeared in Sweden, the Netherlands, and Italy.

The means and standard deviations of the total test scores and the amount of home resources for learning, by mother's language status appear in Table 3. The highest mean test scores for children whose mothers spoke the language of the test were found in Ireland, the Russian Federation, and Sweden, while the lowest scores for this group were in Romania, Norway, and Lithuania. Among

TABLE 3
Mean (SD) Test and Home Resources Scores for Children Whose Mothers Speak the Language of the Test and Those Whose Mothers Do Not

Nation (N)	Mean (SD) Score for Children Whose Mothers Speak Language of Test	Mean (SD) Score for Children Whose Mothers Do Not Speak Language of Test	Mean (SD) Home Resources Mom Speaks Language	Mean (SD) Home Resources Mom Does Not Speak Language
Ireland (401)	10.3 (2.3)	9.7 (2.7)	11.1 (1.6)	10.5 (1.7)
Italy (506)	9.7 (2.2)	9.2 (2.8)	10.0 (1.5)	9.4 (1.5)
Lithuania (700)	8.8 (2.5)	9.0 (2.0)	10.1 (1.5)	10.0 (1.9)
Netherlands (395)	9.7 (2.2)	9.0 (2.2)	11.0 (1.5)	10.0 (2.1)
Norway (465)	8.7 (2.4)	8.4 (2.4)	11.7 (1.3)	11.4 (1.8)
Poland (818)	9.4 (2.5)	9.0 (1.4)	10.3 (1.8)	11.8 (1.1)
Portugal (597)	9.6 (2.4)	10.1 (2.1)	10.2 (1.8)	10.4 (0.7)
Romania (715)	8.5 (3.0)	6.5 (2.4)	9.3 (2.0)	7.8 (2.7)
Russian Federation (713)	10.0 (2.3)	8.9 (2.9)	10.7 (1.3)	9.7 (1.3)
Slovak Republic (933)	9.2 (2.4)	7.8 (2.9)	10.2 (1.6)	8.5 (2.4)
Sweden (555)	10.0 (2.1)	8.8 (3.0)	11.7 (1.6)	10.5 (1.6)

those children whose mothers did not speak the testing language, the highest scores were in Portugal and Ireland, while the lowest means were among examinees from Romania. In terms of home resources for learning, the highest means appeared in Sweden, Norway, Ireland, and the Netherlands for those individuals whose mothers spoke the testing language, and Poland, Norway, Sweden, and Ireland for those whose mothers did not speak the language. The lowest mean resources were in Romania for both groups of examinees. Statistical differences among these variables were not assessed as this was not the focus of the analyses. Instead, this information provides a more broad description of the sample.

Model-Based Recursive Partitioning

The results of the MBRP analyses for each of the 13 test items appear in Table 4. Across all items and all terminal nodes for the MBRP trees, the total test score had a statistically significant positive relationship with performance on the item. This result was expected (i.e., examinees who were better readers were more likely to answer each item correctly), and will not be discussed. For item 1, the MBRP found three distinct logistic regression models based on nation. The first model applied to the nations of Ireland, Italy, Lithuania, the Netherlands, Poland, Portugal, and

TABLE 4
Model-Based Recursive Partitioning Coefficients for Mother's Language and Educational Resources by Item and Nation

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
2	372, 380, 528, 616, 620, 642, 643, 703	0.42*	−0.37	**0.05***
	440, 578, 752	0.37*	0.30	0.02
3	380, 528, 578, 620, 643	0.44*	−0.25	**0.07***
	440, 642	0.45*	−**0.94***	**0.13***
	703	0.37*	0.64	−0.03
	372, 616	0.50*	−**1.49***	0.01
	752	0.72*	−0.46	−0.06
4	528, 578, 616, 620, 643	0.54*	0.14	0.01
	372, 380, 440, 642, 703, 752	0.54*	0.37	0.04
5	372, 528	0.60*	0.36	**0.19***
	440, 616, 620, 642, 703	0.52*	−0.56	0.01
	643	0.50*	**1.52***	0.17
	380, 578, 752	0.53*	0.38	0.05
6	642	0.50*	0.83	−0.01
	372, 578, 620, 643	0.59*	−0.36	−0.06
	380, 440, 703, 752	0.50*	0.15	−0.02
	528, 616	0.53*	−0.65	0.03
7	440	0.41*	−0.32	−0.05
	578, 703	0.69*	0.38	0.01
	642	0.44*	−0.25	0.06
	372, 380, 528, 616, 620, 643, 752	0.70*	0.10	0.02
8	372, 528, 578, 620	0.65*	−0.38	−0.04
	380, 440, 616, 642, 703	0.59*	0.06	**0.07***
	643, 752	0.75*	−0.11	−0.01
9	372, 380, 616, 642, 703	0.71*	0.01	0.02
	440, 528, 578, 620, 643, 752	0.67*	−0.22	**0.08***
10	440, 528, 643	0.45*	0.16	0.06
	372, 380, 578, 616, 620, 642, 703	0.53*	0.18	0.05
	752	0.49*	−**1.12***	−0.04
11	578	0.36*	−0.47	−0.03
	616	0.39*	−13.43	−0.01
	620	0.39*	−15.08	0.06
	380, 440, 642, 643, 703	0.36*	0.22	−0.01
	372, 528	0.48*	**0.72***	−0.02
	752	0.52*	−0.27	−0.06
12	642	0.80*	0.03	0.02
	372, 620, 643, 703	0.79*	0.06	0.01
	528, 578	0.78*	−0.50	−0.00
	380, 440, 616, 752	0.90*	−**0.61***	**0.08***

(Continued on next page)

TABLE 4
Model-Based Recursive Partitioning Coefficients for Mother's Language and Educational Resources by Item and Nation *(Continued)*

Item	Nations[1]	Total Score	Mother's Language	Educational Resources
13	440, 616, 620, 642, 703	0.43*	0.04	0.05
	380, 528	0.59*	0.25	−0.03
	372, 578, 643, 752	0.51*	−0.15	0.02

[1] Numeric codes for nations included in the study: 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 ($\alpha = 0.05$).

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

the Russian Federation. For this group, there were not statistically significant relationships between mother's language and available educational resources, when controlling for total score. In other words, uniform DIF was not present based on either of these variables for this first group of nations. The second set of nations partitioned by MBRP included Norway, Romania, and the Slovak Republic. For these nations there was a statistically significant negative relationship between mother's language status and performance on the item when controlling for total test score and available educational resources. This latter variable was not significantly related to item performance, however. This result indicates that examinees whose mothers did not speak the language of the test responded correctly to item 1 at a significantly lower rate compared to those examinees whose mothers did speak the test language, after conditioning on both overall reading ability in the form of the total test score, and the amount of educational resources available in the home. That is, uniform DIF was present. The final terminal node in this tree included Sweden, for which there was not a statistically significant relationship between mother's language status and available educational resources in the home. One possible reason for Sweden being placed in a unique terminal node is that the relationship between the total score and performance on the item was somewhat lower than was the case for the other nations, and the magnitude of the coefficient for mother's language, although not statistically significant, was between those of the other two nation sets.

Figure 1 includes the MBRP tree for item 1. At the top of the figure are the splits for the tree. Within each node is the *p*-value for the test statistic of the null hypothesis that a split should be made. Another way to consider this test is that it assesses the null hypothesis of parameter stability across nations within the node. Superimposed on the line coming down from each node are the PIRLS code numbers for the nations associated with the split. As an example, in the first split,

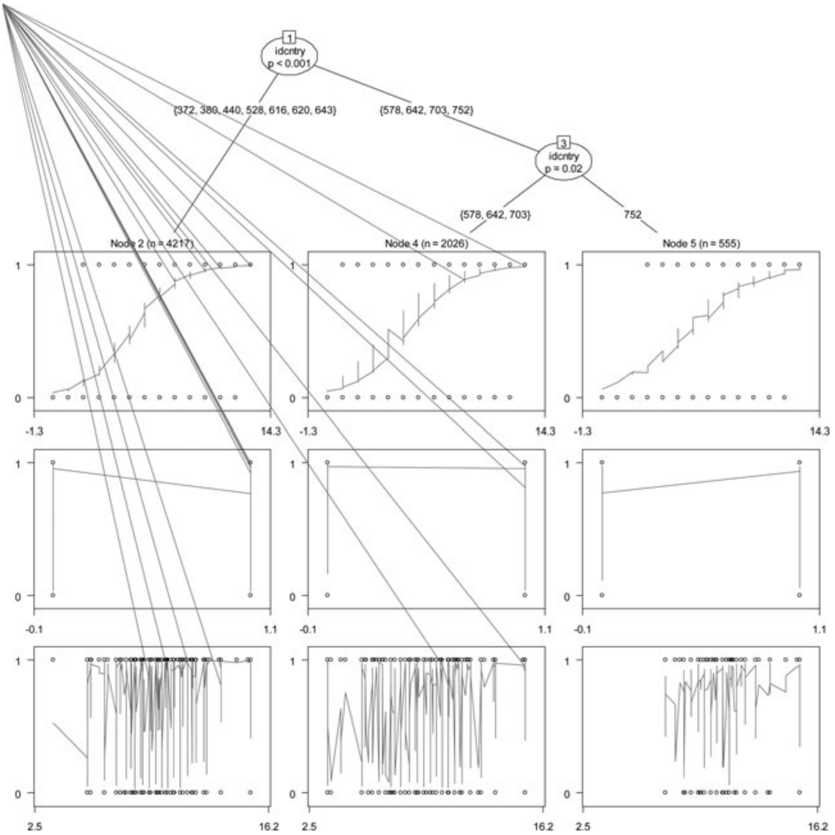

FIGURE 1
Model-based recursive partitioning tree for item 1. *Note*. The top of the figure displays the tree. Within each node is the *p*-value for testing the null hypothesis that a split should be made. Panel graphs appear below the tree showing relationships between the independent variables and outcome (correct item response or not) for each terminal node. Note that these results represent the unconditional relationships and not those conditioned on the full model.

nations 372, 380, 440, 528, 616, 620, and 643 were all placed in the node to the left, and nations 578, 642, 703, and 752 were placed in the node to the right. Below the tree are panel graphs showing the relationships between each of the independent variables and the outcome (correct item response or not) for each of the terminal nodes. Each of the three terminal nodes has associated with it three different plots, one for each independent variable in the LR model (score, mother's language, and educational resources). The first row of plots represents the relationship between score on the assessment (*x*-axis), and probability of a correct response to item 1 (*y*-axis). Thus, we can see that for each terminal node, the probability of a correct response to item 1 increases concomitantly with increases in the scale score. The second row of graphs in this figure displays the relationship between mother's language (*x*-axis) and probability of a correct item 1 response on the *y*-axis. From these results, it appears that for each terminal node, the probability of a correct item response was high, regardless of mother's language. Finally, the third row of plots includes the available educational resources score (*x*-axis) and the probability of a correct item response (*y*-axis). These results suggest that for this item there was not a strong relationship between the two variables for any of the nodes. When interpreting these values it is important to note that the unconditional relationships are being graphed here, while the results presented in Table 4 are conditional in nature. Thus, patterns reported in the table may differ somewhat from those apparent in the graph.

With respect to item 2, MBRP identified two different patterns of relationships for logistic regression. The first terminal node included Ireland, Italy, the Netherlands, Poland, Portugal, Romania, the Russian Federation, and the Slovak Republic. The other node included Lithuania, Norway, and Sweden. The essential difference between these two nation sets was that for the first group there was a statistically significant positive relationship between available home resources and performance on the item (examinees with more resources had a higher probability of providing a correct item response), when conditioning on overall reading ability and mother's language status, whereas for the second nation set there was no such relationship.

For item 3, five separate terminal nodes were identified by MBRP. The first included Italy, the Netherlands, Norway, Portugal, and the Russian Federation. For these nations, when controlling for overall reading ability and mother's language status, there was a statistically significant positive relationship between available educational resources and performance on the item. The second terminal node consisted of Lithuania and Romania, and was associated with a significant negative relationship between mother's language status and performance on the item, and a positive relationship between available educational resources and item performance, when conditioning on the overall reading ability. The third terminal node included the Slovak Republic, for which no DIF was found in terms of either mother's language or educational resources. The fourth terminal node was made up of Ireland and Poland, for which there was significant DIF based on mother's language status (children whose mothers did not speak the testing language performed worse on the item compared to those whose mothers did, after conditioning on overall reading ability and educational resources). Finally, the fifth terminal node included Sweden, for which no DIF was identified.

No DIF was identified for items 4, 6, 7, or 13. For item 5, MBRP identified four terminal nodes, with nodes 1 (Ireland and the Netherlands), and 3 (Russian

Federation) displaying DIF, and nodes 2 (Lithuania, Poland, Portugal, Romania, and the Slovak Republic), and 4 (Italy, Norway, and Sweden) showing no DIF. For nations in node 1, there was a statistically significant positive relationship between available educational resources and the likelihood of a correct item response, while for node 3 examinees whose mothers did not speak the language of the test had a significantly higher probability of providing a correct item response than those whose mothers did speak the language. Again, these significant DIF results were found after conditioning on overall reading ability and the other variable (mother's language or educational resources). With regard to item 8, MBRP identified three terminal nodes. Node 1 (Ireland, the Netherlands, Norway, and Portugal) and node 3 (the Russian Federation and Sweden) did not exhibit DIF in terms of mother's language status or available educational resources. Node 2 (Italy, Lithuania, Poland, Romania, and the Slovak Republic) did display DIF based on available educational resources, such that having more educational resources in the home was associated with a higher likelihood of a correct item response, after controlling for reading ability and mother's language status.

For item 9, two terminal nodes were found such that node 1 (Ireland, Italy, Poland, Romania, and the Slovak Republic) exhibited no DIF, and node 2 (Lithuania, the Netherlands, Norway, Portugal, the Russian Federation, and Sweden) was found to have a statistically significant positive relationship between available educational resources in the home, and performance on the item. For item 10, three partitions were identified, with neither node 1 (Lithuania, the Netherlands, the Russian Federation), nor node 2 (Ireland, Italy, Poland, Portugal, Romania, and the Slovak Republic) displaying DIF for either mother's language status or available educational resources. DIF was found for mother's language status for node 3 (Sweden), whereby examinees whose mothers did not speak the testing language performed worse on the item than those whose mothers did, after controlling for reading ability and available educational resources.

With respect to item 11, MBRP identified six nodes. Of these, only node 5 (Ireland and the Netherlands) displayed DIF. In this case, individuals whose mothers did not speak the testing language had a higher likelihood of getting the item correct than those whose mothers did speak the testing language, after conditioning on the total test score and the available educational resources. For nodes 1 (Norway), 2 (Poland), 3 (Portugal), 4 (Italy, Lithuania, Romania, the Russian Federation, and the Slovak Republic), and 6 (Sweden) no DIF was identified. Finally, for item 12, a total of four terminal nodes were identified. In nodes 1 (the Russian Federation), 2 (Ireland, Portugal, Romania, and the Slovak Republic), and 3 (the Netherlands and Norway) no DIF was found, although there did appear to be differences in the strength of relationship between the total test score and the item response. For node 4 (Italy, Lithuania, Poland, and Sweden), there was significant DIF associated with both mother's language status and available educational resources in the home. For the former variable, examinees whose mothers did not speak the testing language

TABLE 5
Items for which Uniform DIF Was Identified for Mother's Language Status and Educational Resources in the Home Index

Nation	Mother's Language Status	Educational Resources
Ireland	3, 11	2, 5
Italy	12	2, 3, 8, 12
Lithuania	3, 12	3, 8, 9, 12
Netherlands	11	2, 3, 5, 9
Norway	1	3, 9
Poland	3, 12	2, 8, 12
Portugal	2	3, 9
Romania	1, 3	2, 3, 8
Russian Federation	5	2, 3, 9
Slovak Republic	1	2, 8
Sweden	10, 12	9, 12

had a lower likelihood of providing a correct item response, and those with more educational resources in the home were more likely to respond correctly.

Table 5 summarizes the items for which some type of uniform DIF was found, by the nations included in this study. From this table we can see that Norway, Portugal, and the Slovak Republic exhibited the fewest instances of DIF, with three cases each. In addition, DIF was found for three separate items for Sweden,

TABLE 6
DIF Coefficients for Multilevel Models by Item

Item	Score	Mother's Language	Educational Resources	Intraclass Correlation
1	0.07*	−0.02	0.00+	0.15
2	0.06*	0.00	−0.00	0.05
3	0.07*	−0.03	−0.01*+	0.19
4	0.05*	0.03	−0.00	0.10
5	0.07*	0.04+	0.00	0.16
6	0.09*	−0.00	−0.01	0.10
7	0.09*	0.01	0.00	0.07
8	0.10*	−0.01	0.00	0.08
9	0.10*	0.00	0.00	0.12
10	0.05*	0.01+	−0.00	0.13
11	0.08*	0.01+	−0.00	0.19
12	0.08*	−0.05*	0.01	0.10
13	0.08*	−0.00	0.00	0.11

*Statistically significant fixed effect.
+Statistically significant random effect.

although a total of four such cases were in evidence because DIF was found for both mother's language status and available educational resources for item 12. The nations exhibiting the most instances of DIF were Lithuania, Italy, and Poland. In addition, results in Table 5 reveal that DIF was more commonly associated with the availability of educational resources in the home than as a function of mother's language status, except in Ireland where they were equally likely to be present.

Multilevel Modeling with LR

Multilevel model results for each of the 13 items appear in Table 6. The intraclass correlation (ICC) values appear in the last column of the table, and indicate the degree of relationship in the dependent variable within the nations. Larger values suggest stronger such relationships. For the nations used in this study, the ICC values range from 0.07 for item 7 to 0.19 for items 3 and 11. The other columns in Table 6 contain the model coefficients for each variable included in the DIF analysis, with an indicator as to whether it was statistically significant, and whether the random coefficient effect was statistically significant as well. To recall, a significant random coefficient term would indicate that the coefficient values differed across nations. Based on these results, uniform DIF associated with mother's language status was found for item 12, such that examinees whose mothers did not speak the language of the test had a lower probability of a correct item response than those whose mothers did speak the testing language. With respect to educational resources in the home, uniform DIF was found to be present for item 3, whereby individuals with more educational resources in the home had a lower probability of obtaining a correct item response, conditioning on total reading ability. In addition to these two main effects for mother's language and educational resources in the home, there were also statistically significant random coefficient terms for both variables as well. For mother's language, the random effects for coefficient for items 5, 10, and 11 were found to be significant. These results indicate that there were significant differences in the relationship between mother's language and performance on the item across the nations, after conditioning on the total score and the amount of educational resources available in the home. Similarly, the random effects for the coefficient of educational resources were statistically significant for items 1 and 3, indicating that there were differences in the nature of educational resources based DIF for these items across nations.

DISCUSSION

Numerous statistical methods have been developed and tested for assessing the type (uniform or nonuniform) and degree of DIF present for individual items, and for bundles of items (Millsap, 2011). Each of the widely used methods focuses on comparing item performance between a small number of groups conditioning on some measure of overall ability, typically the total score on the instrument. Regardless of differences among them, each of these methods makes a tacit assumption that outside of the grouping variable of interest, respondents are homogeneous with regard to their test and item performances. Thus, for example, when native language–based DIF is assessed for individuals from different states within the United States, as with NAEP, or different nations across the world, such as in PIRLS, TIMMS, or PISA, using a common DIF statistical method such as logistic regression, it is assumed that the only salient variable of interest is native language status. No other variables are typically integrated into the models. This issue is of particular import as interest in large multistate and multinational databases is increasing.

Given the additional issues of assessing DIF in situations where complex sampling designs are used, the goals of this study were twofold. First, the prevalence of uniform DIF associated with mother's language relative to the testing language was of interest, as it may reflect the degree to which large standardized testing efforts fairly represent examinee ability in an important skill such as reading. Because reading serves as a cornerstone for nearly all other academic tasks (Duncan et al., 2007; Sparks, Patton, & Murdoch, 2013), the accurate assessment of this construct is crucial. Children who have difficulty reading benefit from early identification, as remedial attention can occur immediately (Little et al., 2012; Menzies, Mahdavi, & Lewis, 2008). Thus, they have a lower risk of falling behind their peers in reading and then potentially other academic constructs (e.g., mathematics). Therefore, reading assessments must produce accurate scores for all examinees.

To this end, the presence of uniform DIF was assessed in the current study, conditioning on both the score on the reading assessment, and the available educational resources in the home. This latter variable was included to provide statistical control and contextualization of the DIF results, given its importance in terms of reading achievement (Herbers et al., 2012). In other words, it was of interest whether, when controlling for available educational resources, DIF would be found for mother's spoken language in the home. Explanation as to the causes of this DIF would involve item content review by experts. The second goal of this study was to explore use of MBRP as a tool for DIF assessment, allowing for differences in the nature (and even the presence) of DIF across large collections of examinees, such as nations or states. The approach employed was based on MBRP in which separate LR models were fit to subgroups of the total sample, based on model coefficient differences across nations. Of particular interest was the degree to which mother's language-based DIF was present in the PIRLS reading assessment, across 11 European countries. By relying on the MBRP approach, we clearly delineated differences in the presence and magnitude of DIF based on maternal language status and available educational resources in the home. In this way, we did not make the a priori assumption that DIF was consistent across nations, but rather allowed for the possibility that some subgroups of nations did share similar patterns of DIF, while other subgroups of nations had different such patterns.

Our results demonstrated that whether the mother spoke the language of the test in the home was related to the presence of uniform DIF on the PIRLS reading assessment even when controlling for available educational resources in the home. With one exception, the presence of DIF was associated with lower scores for examinees whose mothers did not speak the language of the test, even when they were matched on reading ability with examinees whose mothers did speak the testing language. Thus, it would appear that within some nations, certain items present a disadvantageous linguistic load for examinees whose mothers did not speak the testing language in the home, even when accounting for their overall reading ability and the amount of educational resources in the home. The amount of available educational resources in the home were also associated with uniform DIF for several items, such that those with more educational resources performed better on some items, even when matched with individuals on overall reading ability and the same maternal language status. Much prior research has demonstrated that in general SES is associated with performance on academic achievement tests. The current study extends this work by demonstrating that the resources in the home, in this case those specific to education and learning, are associated with differential item performance even for individuals who have equal levels of the latent trait being measured and the same maternal language status. Furthermore, DIF was also found to be present for students whose mothers did not speak the testing language in the home. Finally, these patterns were not consistent across the nations examined. MBRP identified subgroups of nations for the various items for which DIF was present, and others for which it was not. Taken together, these results suggest that the relationship of maternal language status, student economic standing, and achievement is more complex than might have been assumed based on prior research. Furthermore, these results demonstrate that the presence of such DIF was not consistent across nations. For a given item, DIF based on the mother's language and the available educational resources were present in some countries but not in others.

Some important insights into the DIF results reported here can be gleaned from prior research regarding the issue of DIF with respect to differences in the native language of examinees. In particular, there is evidence that language-based DIF is more likely to occur when the language of the test and the native language of the examinee are more dissimilar linguistically (Grisay & Monseur, 2007; Kim, 2001). Even more specific to the European sample studied here, using multidimensional scaling Robin, Sireci, and Hambleton (2003) found that a smaller proportion of items on an international credentialing exam were found to exhibit DIF among European examinees when the non-native language spoken was another European language than when it was from a non-European language group.

Indeed, even within Europe, differences in item difficulty have been found across languages, suggesting that different linguistic structures even for relatively similar languages may contribute to the presence of DIF (Grisay, de Jong, Gebhardt, Berezner, & Halleux-Monseur, 2007). Drummond (2011) found that for a high-stakes scholarship test in the Kyrgyz Republic, cross-national DIF was identified for a number of items, but that some item types (e.g., those measuring reading comprehension) were less susceptible to DIF than were items measuring analogy and sentence completion.

These prior research findings can be further considered in light of immigration patterns for the nations included in the current study, which differed quite widely. As reported by the European Union (European Commission, 2013), for example, the largest group to immigrate to Sweden was from Finland (11.1%), whereas for Italy the largest group of immigrants was Romanian (17.6%), and for the Netherlands the largest group came from Turkey (10.2%). There is not room in the current manuscript for an exhaustive list of the immigration patterns for these nations. However, even a cursory examination such as this highlights the great variation in immigration patterns across the nations included in the study. Taking the findings from prior research described previously together with the different immigration patterns and the likely resulting differences in languages spoken in the home, it is very possible that some of the differences in cross-national DIF results reported here may be due to cross-national differences in mother's language when that language is not the language of the test, rather than strictly a function of less knowledge and comfort with the language of the test by examinees.

Some of the DIF results presented for MBRP might also be interpreted in light of technical characteristics of the items themselves. In particular, item 12 was found to exhibit DIF quite frequently, and that it also was the most difficult item in the assessment. A number of authors have shown that a positive relationship exists between item difficulty and findings of DIF (e.g., Freedle, 2003; Kulick & Hu, 1989; Scherbaum & Goldstein, 2008). Santelices and Wilson (2012) further investigated the relationship between item difficulty and findings of DIF in order to explore the argument by Dorans and Zeller (2004), among others, that such a relationship might be primarily due to a statistical artifact. However, Santelices and Wilson found that this was not likely the case, and that indeed such a relationship does exist, and that it is independent of the statistical methodology chosen to investigate DIF. Indeed, Freedle (2003) argued that one possible cause for such findings of DIF is the difference in the language spoken in the home and in school, whether these languages be variants of the test language or completely different, as was the case in this study. Thus, the current study results appear to provide further evidence for Freedle's hypothesis, and also suggest avenues for future research in which the relationship between item difficulty and language status is more directly investigated in the international context.

Comparison of MBRP and Multilevel LR

In addition to MBRP, we also examined the issue of DIF with respect to mother's language in the home, and amount of available educational resources using a more traditional method, the MLR model. This approach yielded somewhat different results from the MBRP. Specifically, statistically significant random coefficient effects were found for items 5, 10, and 11, with respect to Mother's Language, and items 1 and 3 with respect to Educational Resources. Substantively these results mean that the multilevel model identified significant cross-national differences with regard to uniform DIF associated with mother's language and educational resources for these items. In addition, uniform DIF across all nations with respect to item mother's language was only found to be present for item 12, and for item 3 for educational resources.

When comparing the results of the two methods with one another different patterns clearly emerge. While it is not possible to say definitively which approach (MBRP or MLR) is correct, given that we do not have access to the population itself, it is possible to compare results of the two approaches in an attempt to glean an understanding as to why they might provide different answers to the question of whether DIF was present. In particular, a review of the DIF results presented in Table 4 shows that the coefficients estimated in the context of MBRP differed a great deal across clusters of nations. Thus, it is entirely possible that when these separate models are combined into a single multilevel model, the differences in cross-national effects may cancel one another out, leading to the conclusion reached by MLR that no significant DIF was present overall (i.e., there was no significant fixed effect for mother's language or educational resources for most items). And indeed, a comparison of results in Tables 3 and 4 would appear to support this hypothesis, at least descriptively.

Consider the educational resources coefficients on item 3 for the nation subgroups identified by MBRP (Table 3). These values were statistically significant and positive for two of the nation subsets, which together comprised 7 of the 11 nations in the study, and just over 60% of the total sample of examinees. Thus, it is not surprising given this relative homogeneity of results indicating DIF in a particular direction that an overall finding of DIF across nations might be found. Similarly, for item 12, which was found to have a significant fixed effect with respect to mother's language, the nation specific estimates associated with MBRP (Table 3) are of interest. In this case, 6 of the 11 nations belonged to subgroups that together comprised 47.3% of the sample of examinees. While this percentage is slightly less than half, the size of the coefficients for mother's language for these two nation subgroups was very large when compared to the values for the other two nation subsets, both of which were near 0. Thus, again, it would appear that the relative homogeneity of the findings with MBRP are associated with a statistically significant fixed effect for a variable in the MLR model used to test for DIF.

With regard to the random coefficients in the MLR model, significant results were found for items 1, 3, 5, 10, and 11. As noted, the conclusion to be drawn from these results is that the relationships between the main effect (i.e., mother's language or educational resources) and performance on the item differed across nations, after conditioning for the total test score. In contrast, MBRP identified multiple cross-national subgroups for each of the items, indicating that there were differences on at least one of the coefficients (score, mother's language, and educational resources) across nations. In considering the results presented here, at least two possible explanations for this apparent discordance in results may emerge. First, an examination of Table 4 reveals that the ICCs for those items with significant random coefficients in the multilevel LR analysis were larger than was the case for the other items, for which no significant random coefficient effects were found. The mean ICC for items 1, 3, 5, 10, and 11 was 0.16, whereas the mean ICC for the other items was 0.09. Therefore, one possible hypothesis for explaining the lack of statistically significant random effects for the coefficients is that the lower ICC values deflated test statistics, thereby limiting the number of significant results that were obtained. In fact, low ICCs in combination with small samples and small DIF magnitudes has been associated with lower statistical power for multilevel models for DIF detection (French & Finch, 2010 2013).

A second issue to consider is the method by which MBRP determines whether and how to divide the data as it builds the tree, as compared to the test used by multilevel models to determine whether a random coefficient is statistically significant. As noted, the MBRP methodology examines every possible way in which nations could be divided, and for each of these assesses the model parameters in order to determine whether there exists sufficient evidence to divide the sample. Thus, it is possible that collectively, across all model parameters, there is sufficient instability so as to warrant the recursive partitioning algorithm making a split. In contrast, the test for a specific random coefficients term in multilevel LR is only assessing that coefficient and whether there is sufficient evidence that it differs across nations (in this application) to suggest that multiple such values are present in the population. While it is true that the estimates for the random coefficients effects are conditioned on other variables in the model, the test itself is only examining a single such effect (e.g., mother's language). Again, MBRP is assessing whether collectively across parameters models differ across nations. Given these differences, it is certainly possible that the two methodologies could yield somewhat different results for the same set of data.

When considering the results of this study as a whole in regards to which methodology might be most useful under what circumstances, the researcher should consider a few important issues. First, both MBRP and MLR can yield useful and important information in DIF investigations involving groups or clusters of examinees, such as nations. MBRP excels at providing the researcher with explicit groups of such clusters, and parameter estimates for each such group so that the researcher can see immediately how the nature of DIF might vary across clusters. In addition, the researcher will know that there are significant differences in the parameter sets for different cluster groups, although they will not know which of the parameters differ across sets. Therefore, if the primary interest of the researcher is to identify subgroups of clusters, such as nations, states, or schools, for which the pattern of DIF is similar, MBRP may be the optimal analysis. On the other hand, the MLR provides an explicit test for whether there are differences in the nature of DIF, in the form of tests for the random coefficients term in the model. This approach will not yield specific parameter estimates for each cluster should the random coefficient be found significant, requiring the researcher interested in such values to fit separate LR models to each cluster (nation) separately. Thus, for researchers interested in measuring an overall DIF effect for a specific variable (e.g., mother's language), and who merely need to account for the clustering of examinees but have no inherent interest in the clusters themselves, MLR may be preferable as an analytic strategy. However, please note that this approach does have the potential to miss subtle and potentially informative differences in model results across clusters of nations that would be potentially detectable by MBRP.

While the analysis selected for comparison to MBRP in this study was multilevel IRT, it should also be noted that MBRP also has some commonalities in purpose with the explanatory IRT models described by De Boeck and Wilson (2004). The goal of those models was to incorporate additional information about the respondents in an effort to better understand the mechanisms underlying item responses. Such explanatory models make use of multilevel IRT techniques as well as multilevel logistic and other regression approaches that can provide researchers with additional information about relationships among external variables (e.g., respondent SES or maternal language) and responses to items. In this regard, MBRP and explanatory IRT can both be used to gain greater insights into item response patterns, although they obviously differ methodologically. In the context of explanatory IRT, maternal language and family SES could be included as level-1 predictor variables in a multilevel IRT model, and relationships between these variables and the item response parameters would then be estimated. If the researcher were interested in the impact of nation, it would need to be included as a level-2 predictor variable, and differences in the impact of maternal language and SES on the item responses, across nations, could be modeled using cross-level interactions of these variables. As was demonstrated, MBRP incorporates such interactions somewhat differently, by using nation as a partitioning variable and identifying groupings of countries for which the relationship between item responses and maternal language and SES were similar. Nonetheless, there is certainly some correspondence between the work of De Boeck and Wilson, and MBRP.

Directions for Future Research and Study Limitations

Working under the assumption that either variable studied here is associated with a consistent type of DIF across national borders may not be accurate. Rather, the nature of uniform DIF with respect to these items appears to vary from nation to nation. Further work in this area should focus on developing a deeper understanding of why DIF for maternal language in the home and available educational resources might differ across nations. Such future research would potentially include an educational policy review component, expert review of item content associated with DIF in each of the nations with a particular focus on the languages spoken by the mothers and how these languages differ linguistically from the language of the test, and more comprehensive sampling inclusive of acculturation variables in which detailed interviews of parents and observations in the home and classroom are combined to provide a more complete picture of how language, economics, and acculturation interact to influence achievement across nations. We recognize that such a large-scale interview effort would need to be conducted under the aegis of the PIRLS program in order to meaningfully involve all of the nations participating in the testing program. In conjunction with this recommendation, future work should also examine the extent to which these DIF findings may be due to differences in mother's language across non-native speakers in the different nations. In addition, future research should ascertain whether cross-national differences in the presence and nature of DIF are seen outside of Europe. An assessment of nonuniform DIF would also be important in order to determine the extent to which specific items might differ in terms of differentiating examinees with different reading abilities, and how such differences might themselves differ across nations.

Another limitation of the current study is with respect to the sample size differences between the children whose mothers spoke the language of the test, and those whose mothers did not. For all nations, the large majority of individuals came from homes where the mother spoke the language in which the test was administered. This sample size imbalance raises two issues that might impact the interpretability of the study results, and about which researchers using MBRP should be aware. First, when the sample size within a node is very small, the resulting model parameter estimates (e.g., slopes) and their associated standard errors may not be stable. Therefore, when interpreting these estimates it is crucial to consider how many individuals are contained within each node. When that number is small, the estimates must be interpreted with greater caution. A second consideration that needs to be made when the sample sizes are very unequal is with respect to the actual grouping of individuals into nodes. Research with CART has shown that very unequal sample sizes can result in classification inaccuracy because members of smaller partitioning group(s) tend to be placed with those in larger group(s) (Holden, Finch, & Kelly, 2011). Thus, in the case of CART, group separation can be compromised with very unequal sample sizes. Research has not yet been published considering whether this problem is present with MBRP as well, but given its methodological similarity to CART, the problem of misclassified node membership must be considered when interpreting MBRP results. In addition to these marked differences in sample sizes, another issue to consider in this study is the relatively small number of items that make up the Flowers assessment. In this case, the 13 items in the scale might be on the lower end of what would be desired for obtaining appropriate IRT parameter estimates (de Ayala, 2009). Given that a natural follow-up to the significant uniform DIF results (although not one used here) would be the estimation of difficulty for each item by terminal node, the fact that there were 13 items in the scale would need to be considered when interpreting the values.

Finally, it should be noted that scale purification was not conducted in this study given the current debate of its usefulness to DIF detection accuracy (Magis & Facon, 2012). In the context of MBRP purification becomes a difficult issue to address because of the differences in the presence of DIF across the levels of the partitioning variable (e.g., nation). For example, in the results presented for a given item, some nodes displayed DIF with regard to the mother's language spoken in the home, while for other nodes no DIF was found. Given that this information will not be known ahead of time, item purification would become a problematic task. However, it is also important to acknowledge that particularly when using logistic regression for DIF detection, item purification can help to limit the number of false positive findings (French & Maller, 2007). Given the difficulty in applying standard item purification methods to the MBRP approach, coupled with its potential utility, further work needs to be done for developing a more flexible purification technique that can be applied in this context.

REFERENCES

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational assessment*, *8*(3), 231–257. doi:10.1207/S15326977EA0803_02.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Journal of Measurement in Education*, *14*(3), 219–234.

American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*. Washington, DC: AERA Publications.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Albano, A. D., & Rodriguez, M. C. (2013). Examining differential math performance by gender and opportunity to learn. *Educational and Psychological Measurement*, *73*(5), 836–856.

Balluerka, N., Gorostiaga, A., Gómez-Benito, J. & Hidalgo, M. D. (2010). Use of multilevel logistic regression to identify the causes of differential item functioning. *Psicothema*, *22*(4), 1018–1025.

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

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oakes, CA: Sage.

Chan, K.-Y., & Loh, W.-Y. (2004). LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics*, *13*, 826–852.

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing*, *6*, 57–79

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement*, *33*, 453–464.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133–148.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: The Guilford Press.

de Ayala, R. J., Kim, S-H., Stapleton, L. M., & Dayton, C. M. (2003). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, *2*, 243–276.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York, NY: Springer.

Dorans, N., & Zeller, K. (2004, July). *Examining Freedle's claims and his proposed solution: dated data, inappropriate measurement, and incorrect and unfair scoring (Report No. RR-04-26)*. Princeton, NJ: Educational Testing Service.

Drummond, T. W. (2011). *Predicting differential item functioning in cross-lingual testing: the case of a high stakes test in the Kyrgyz Republic*. Doctoral dissertation. Retrieved from ProQuest LLC.

Duncan, G. J, Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., et al. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446.

European Commission. (2013). *Statistics explained: main countries of citizenship and birth of the foreign-born population, 2013*. Brussels, Belgium: Author.

Freedle, R. (2003). Correcting the SAT's ethnic and social-class bias: a method for reestimating SAT scores. *Harvard Educational Review*, *73*, 1–43.

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, *47*(3),299–317.

French, B. F., & Finch, W. H. (2013). Extensions of Mantel-Haenszel for multilevel DIF detection. *Educational and Psychological Measurement*, *73*, 648–671.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for DIF detection. *Educational and Psychological Measurement*, *67*, 373–393.

Gama, J. (2004). *Functional Trees. Machine Learning*, *55*, 219–250.

Glas, C., & Jehangir, K. (2014). Modeling country-specific differential item functioning. In L. Rutkowski, M. von Davier, and D. Rutkowski (Eds.), *Handbook of international large-scale assessment: background, technical issues, and methods of data analysis*. Boca Raton, FL: CRC Press.

Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, *8*(3), 249–266.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation*, *33*(1), 69–86.

Han, W. J. (2012). Bilingualism and academic achievement. *Child Development*, *83*, 300–321.

Heath, A. F., Rothon, C., & Kipli, E. (2008). The second generation in Western Europe: education, unemployment, and occupational attainment. *Annual Review of Sociology*, *34*, 211–235.

Herbers, J. L., Cutuli, J. J., Supkoff, L. M., Heistad, D., Chan, C.K., Hinz, E., & Masten, A. S. (2012). Early reading skills and academic trajectories of students facing poverty, homelessness, and high residential mobility, *Educational Researcher*, *41*, 366–374.

Holden, J. E., Finch, W. H., & Kelly, K. (2011). A comparison of two-group classification methods. *Educational and Psychological Measurement*, *71*, 870–901.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*, 651–674.

International Association for the Evaluation of Educational Achievement. (2011). *Progress in International Reading Literacy Study*. Boston: Author.

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

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18*, 89–114.

Koo, J., Becker, B. J., & Kim, Y. S. (2014). Examining differential item functioning trends for English language learners in a reading test: a meta-analytical approach. *Language Testing*, *3*, 89–109.

Kulick, E., & Hu, P. G. (1989). *Examining the relationship between differential item functioning and item difficulty (College Board Report No. 89-5; ETS RR-89-18)*. New York: College Entrance Examination Board.

Little, M. E., Rawlinson, D., Simmons, D. C., Kim, M., Kwok, O.-M., Hagan-Burke, S., Simmons, L. E., Fogarty, M., Oslund, E., & Coyne, M. D. (2012). A comparison of responsive interventions on kindergarteners' early reading achievement. *Learning Disabilities Research & Practice*, *27*, 189–202. doi:10.1111/j.1540-5826.2012.00366.x.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, *12*, 361–386.

Magis, D., & Facon, B. (2012). Item purification does not always improve DIF detection: a counterexample with Angoff's Delta Plot. *Educational and Psychological Measurement*, *73*(2), 293–311.

Martin, M. O., Mullis, I. V. S., Foy, P., & Arora, A. (2012). Creating and interpreting the TIMSS and PIRLS 2011context questionnaire scales. In M. O. Martin and I. V. S. Mullis (Eds.), *TIMSS and PIRLS methods and procedures*. Boston, MA: International Association for the Evaluation of Educational Achievement.

Menzies, H. M., Mahdavi, J. N., & Lewis, J. L. (2008). Early intervention in reading: from research to practice. *Remedial and Special Education*, *29*(2), 67–77.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257–274.

Oliveri, M. E., Ercikan, K., & Zumbo, B. (2013). Analysis of sources of latent class differential item functioning in international assessments. *International Journal of Testing*, *13*, 272–293.

Organisation for Economic Co-Operation and Development. (2010a). *PISA 2009 results: overcoming social background, equity in learning opportunities and outcomes*. Paris: Organization for Economic Cooperation and Development.

Organisation for Economic Co-Operation and Development. (2010b). *PISA 2009 results: what students know and can do-student performance in reading, mathematics, and science*. Paris: Organization for Economic Cooperation and Development.

Organisation for Economic Co-Operation and Development. (2013). *PISA 2012 assessment and analytical framework: mathematics, reading, science, problem solving and financial literacy*. Paris, France: Author.

Potts, D., & Sammut, C. (2005). Incremental learning of linear model trees. *Machine Learning*, *61*, 5–48.

R Foundation for Statistical Computing. (2013). *R: a language for statistical computing*. Vienna, Austria: Author.

Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, *3*(1), 1–20.

Santelices, M. V., & Wilson, M. (2012). On the relationship between differential item functioning and item difficulty: an issue of methods? Item response theory approach to differential item functioning. *Educational and Psychological Measurement*, *72*(1), 5–36.

Scherbaum, C., & Goldstein, H. (2008). Examining the relationship between race-based differential item functioning and item difficulty. *Educational and Psychological Measurement*, *68*, 537–553.

Sparks, R. L., Patton, J., & Murdoch, A. (2014). Early reading success and its relationship to reading achievement and reading volume: replication of "10 Years Later." *Reading and Writing*, *27*, 189–211.

Su, X., Wang, M., & Fan, J. (2004). Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, *13*, 586–598.

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

Vermunt, J. K. (2008). Multilevel latent variable modeling: an application in education testing. *Austrian Journal of Statistics*, *37*(3), 285–299.

Walker, C. M., & Beretvas, S. N. (2006). An empirical investigation demonstrating the multidimensional DIF paradigm: a cognitive explanation for DIF. *Journal of Educational Measurement*, *38*(2), 147–163.

Yen, W. M. (1984). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement*, *12*, 353–364.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213.

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488–508.

Zeileis. A., Hothorn. T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.

Zeileis, A., Hothorn, T., & Hornik, K. (2014). Party with the mob: model-based recursive partitioning in R. R. Vignette. *R: a language for statistical computing*. Vienna, Austria. Retrieved from http://cran.r-project.org/web/packages/party/vignettes/MOB.pdf.