Studies in Educational Evaluation 56 (2018) 197–204

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Using recursive partitioning Rasch trees to investigate differential item functioning in second language reading tests

Vahid Aryadoust

National Institute of Education, Nanyang Technological University, 1 Nanyang Walk, Singapore

ARTICLE INFO

Keywords:
Differential item functioning
Gender
Grammar
Reading
Recursive partitioning rasch trees
Vocabulary

ABSTRACT

The present study applied recursive partitioning Rasch trees to a large-scale reading comprehension test (n = 1550) to identify sources of DIF. Rasch trees divide the sample by subjecting the data to recursive non-linear partitioning and estimate item difficulty per partition. The variables used in the recursive partitioning of the data were vocabulary and grammar knowledge and gender of the test takers. This generated 11 non-prespecified DIF groups, for which the item difficulty parameters varied significantly. This is grounded within the third generation of DIF analysis and it is argued that DIF induced by the readers' vocabulary and grammar knowledge is not construct-irrelevant. In addition, only 204 (13.16%) test takers who had significantly high grammar scores were affected by gender DIF. This suggests that DIF caused by manifest variables only influences certain subgroups of test takers with specific ability profiles, thus creating a complex network of relationships between construct-relevant and -irrelevant variables.

1. Introduction

Differential item function (DIF) has been conventionally defined as the degree to which test items (dis)advantage test takers based on their group membership (Zumbo, 2007). There are various techniques to investigate DIF in psychological and educational instruments, such as the application of psychometric models (i.e., Rasch measurement and item response theory), all of which aim to explore whether the instrument's functionality is tainted by sources of construct-irrelevant variance (CIV) (Allalouf, Hambleton, & Sireci, 1999; see also Bartolucci, 2007; Bacci et al., 2014, for multidimensional models of DIF analysis). DIF has been widely investigated in reading research in confirmatory and exploratory models. The former approach tests and confirms prespecified hypotheses about the causes of DIF, whereas the latter approach identifies sources of DIF and postulates theories/hypotheses (Aryadoust & Zhang, 2016).

Regardless of the approach, there are two notable issues surrounding DIF analysis techniques in reading assessment research. The first issue is a lack of cognitive theories backing the choice of grouping variables (covariates), which has resulted in the proliferation of DIF analysis across manifest variables such as gender, age, or first language, to name a few. The second issue is that DIF analysis across manifest variables assumes that items that do not exhibit DIF across the specified manifest variables are not confounded by DIF, which is not always an accurate postulation (Tutz & Berger, 2015). Recent methods of DIF analysis, such as the mixture Rasch models, have revealed that items exhibiting no DIF across manifest variables may still be confounded by latent class DIF (Chen & Jiao, 2014; Cohen & Bolt, 2005; Zumbo, 2007; Zumbo, Liu, Wu, Shear, Astivia, & Ark, 2015)—which remains masked in DIF analysis across manifest analysis. These issues impose certain limitations on DIF research, such as difficulty in theoretical justification of the observed DIF and a lack of contribution of DIF studies to theory building in language and educational assessment fields. Finally, DIF analysis, specifically what Zumbo (2007) refers to as the third generation of DIF detection, may not necessarily point to sources of CIV. By contrast, the third generation DIF detection techniques function like ANCOVA (analysis of covariance) and investigate whether test takers differ based on certain covariates.

To address the aforementioned issues, two new approaches to investigating DIF have been developed: the mixture Rasch model (Rost, 1990; von Davier, & Rost, 1995) and recursive partitioning Rasch trees (Strobl, Kopf, & Zeileis, 2015; Zeileis, Strobl, Wickelmaier, & Kopf, 2010). The former approach conflates Rasch measurement and latent class analysis, dividing the sample into several latent classes across which item difficulty varies; next, external variables and/or a post-hoc qualitative analysis of the items are used to identify the distinctive features of the latent classes (Aryadoust & Zhang, 2015). The second approach, which is used in the present study, uses manifest variables (like traditional DIF analysis) and combines classification and regression trees (CART) and Rasch measurement. However, in this approach, the manifest variable is "recursively partitioned", splitting groups of test takers based on the values of the covariate(s) into subgroups where

E-mail address: vahid.aryadoust@nie.edu.sg.

https://doi.org/10.1016/j.stueduc.2018.01.003
Received 24 November 2017; Received in revised form 3 January 2018; Accepted 22 January 2018
Available online 03 February 2018

item difficulty differs (Strobl et al., 2015). The distinct advantage of the Rasch trees in studying reading comprehension is that one can directly examine the effect of important reading components (e.g., lexical and grammatical knowledge) on learners' reading ability without resorting to speculation (like traditional DIF analysis) or through splitting readers into latent classes without considering prerequisite knowledge bases (like the mixture Rasch model). Therefore, the Rasch tree approach lends itself to evaluating reading within the framework of reading theories.

The present study aims to investigate DIF in a high-stakes reading test that is part of a university entrance exam in an [anonymized location]. Based on reading theories such as the construction-integration (CI) model, it is hypothesized that DIF can be induced by readers' grammar and vocabulary knowledge (Kintsch, 1998), but that this DIF is unlikely to cause CIV in the test scores; rather, these theories would suggest that certain test items rely more heavily on certain knowledge bases, and learners whose grammar or vocabulary knowledge bases are more advanced stand a higher chance of answering those items accurately (see Strobl et al., 2015). This study also used manifest variables, as previous DIF research has found a role for some of these variables in causing DIF (see below). By applying the Rasch trees approach, this study investigated whether recursively partitioned manifest variables can reveal DIF patterns in a reading test and, if so, what portion of the observed DIF causes CIV and jeopardizes the validity of the uses and interpretations of the test scores.

2. Literature review

2.1. Reading comprehension

Virtually all theories of reading comprehension stress the importance of certain subskills or knowledge bases in reading, such as vocabulary and grammar (Grabe, 2009; Koda 2005), (meta)cognitive strategies (Paris & Winograd, 1990), world knowledge (Alderson, 2000), and overall language proficiency (Anderson & Freebody, 1981). Vocabulary and grammar knowledge is a significant component that, according to Perfetti and Stafura (2014), provides readers with the essential capacity for processing linguistic input. This knowledge is especially important in bottom-up reading where readers attempt to comprehend the text by decoding it into smaller lexical and grammatical units and then moving "upward" to larger units such as phrases, clauses, and discourse (Kintsch, 1998). Readers also use top-down processes to integrate their world knowledge, attitudes, and experiences with the text (Stanovich & Stanovich, 2006). According to Aryadoust and Zhang (2016), these approaches to reading are congruent with the component-skill view of reading where readers are perceived to possess multiple discrete yet interrelated knowledge bases and cognitive mechanisms, such as vocabulary and grammar knowledge and metacognitive strategies (Grabe, 2009). Research further shows that both bottom-up decoding processes and top-down processes of meaning construction interact during reading comprehension (Golden & Rumelhart, 1993).

Research has shown that vocabulary knowledge is one of the most influential predictor of reading comprehension (Perfetti, Landi, & Oakhill, 2005), with its role varying as a function of the developmental stage and the deepness of the orthography (for a review, see, for example, Florit & Cain, 2011). For example, early research by Laufer (1989) demonstrated that familiarity with the words in the text is essential for readers to efficiently understand the text. Similarly, the Lexical Quality Hypothesis (Perfetti, 2007) predicts that advanced readers benefit from their vast lexical repertoire, which is consistent with empirical research into reading in first and second languages (Qian, 2002). Hu and Nation (2000) also argued that an encounter rate of one unknown word per 20 words in a text can jeopardize readers' comprehension. Recent research has further established the important role of vocabulary in reading, finding correlations between the two ranging from .50 to .82 (Qian, 2002).

Grammatical knowledge can be considered the other fundamental predictor of reading comprehension (Shiotsu & Weir, 2007). In a recent meta-analysis of reading research, Jeon and Yamashita (2014) reported that grammar skills correlated more highly with second language reading (.85) than vocabulary knowledge (.79), testifying to the important role of grammar in comprehension. (Of course, given that the relative weight of the comprehension predictors varies between different populations, the relative significance of lexico-grammatical knowledge in reading would vary across different groups.) According to Purpura (2004), grammatical knowledge is multidivisible and comprises form, meaning, and pragmatic functions. Form includes phonemes and morphemes; meaning includes denotations and connotations; and pragmatic function includes socio-cultural and psychological aspects of language use. These components monitor reading comprehension (Grabe, 2009) and help a reader generate coherence by forming propositions and discourse-level comprehension (Kintsch, 1998). For example, Yano, Long, and Ross (1994) showed that readers' comprehension of texts was significantly enhanced when the texts were made grammatically simpler despite being lexically intact.

Readers' demographic backgrounds may also play a part in success; however, much like variables such as gender, this role is not yet well established. For example, Chiu and McBride-Chang (2006) found that females had more advanced levels of reading proficiency than males, whereas McGeown, Goodwin, Henderson, and Wright (2012) reported no difference. In contrast, Bügel and Buunk (1996) found that boys outperformed girls in reading comprehension. In DIF research, which is reviewed below, gender has been researched rather extensively alongside other variables, such as age or first language, and several studies support its role in DIF.

In sum, the contributions of grammar and vocabulary to the reading process are strongly supported in the extant research. What remains underresearched is whether variation in lexical and grammatical knowledge characterizes groups of readers and influences their performance and if so, whether this influence is linear as represented by the conventional linear methods of data analysis. This query is quite different from the correlational research reported earlier, and seeks to determine whether differences in either vocabulary or grammatical knowledge bases can result in reading test takers' differential test performance. Such uncovered differences would not contribute to CIV, and can help researchers and practitioners identify the major causes of readers' failure on reading tests and provide relevant remedial programs to assist them. According to Rupp (2005), such a research methodology:

> can be used to uncover sets of variables that indicate differential performance generally and to understand the relative strength and weaknesses of different subpopulations (Klieme & Baumert, 2001). Taken a step further, this implies that even without having to believe in the latent variable score as an indicator of true "ability", one can use the scoring model as a filter to carve out groups of examinees that are unique from others with respect to the instrument as a benchmark (Rupp, 2005, p. 92).

This quantitative technique suitable for this approach is described in further details below.

2.2. DIF in reading assessment

DIF analysis in reading comprehension research has conventionally been performed using manifest variables such as gender, first language, and age. Pae (2011) applied Mantel-Haenszel (MH), item response theory (IRT), and linear multiple regression analysis to identify DIF occasioned by gender in reading comprehension. Pae found that many items exhibited DIF, which was likely induced by interactions between gender and item types. Pae argued that DIF items do not necessarily impose unfairness in the tests, but could indicate multidimensionality caused by secondary constructs. Gnaldi and Bacci (2016) also investigated DIF attributed to gender and region, using a multidimensional latent class 2PL (2 parameter logistic) IRT, which identified five latent classes across a battery test of reading, grammar, and mathematics, which they attempted to justify by including "covariates at student and school levels" (p. 53). Cadime, Viana, and Ribeiro (2014) also investigated the effect of region on DIF in a test of reading taken by students from rural and urban areas using logistic regression and MH DIF techniques. They identified 17 (out of 30) non-substantive DIF cases, which they argued did not taint the test results.

In another study, Aryadoust and Zhang (2016) applied a mixture Rasch model to a test of reading comprehension used in Chinese universities. Two latent classes emerged in the analysis, which the researchers subsequently tried to characterize using readers' lexical and grammatical knowledge, (meta)cognitive strategies, and gender. Class 1 had a higher probability of correctly answering reading-in-depth items and had higher vocabulary and grammatical knowledge and general English proficiency. On the other hand, class 2 outperformed class 1 in skimming and scanning items, but had lower vocabulary and grammatical knowledge levels and general English proficiency. Interestingly, gender had a negligible one-to-one association with the performance of the readers, providing further support for previous research where gender and other manifest variables were insufficient to explain DIF (Chen & Jiao, 2014; Hong & Min,2007).

Age is another variable investigated in DIF research. Ownby and Waldrop-Valverde (2013) used nonparametric IRT to investigate whether response format has an influence on older readers in a cloze test of health literacy. They identified 24 (out of 50) DIF cases, which they interpreted as actual DIF rather than any differences resulting from health literacy. Accordingly, the observed DIF was treated as a significant source of CIV, which would jeopardize the uses and interpretations of the test scores.

Finally, Koo, Becker, and Kim (2014) performed meta-analytic DIF analyses with Mantel-Haenszel (MH) on a reading test and the Florida Comprehensive Achievement Test (FCAT). They divided their sample into English language learners (ELLs) and non-ELLs, as well as by gender and ethnicity. They found that items engaging vocabulary knowledge and phraseology in context favoured non-ELLs in grade 3, while items requiring evaluation skills favoured ELLs in grade 10. They further found that Grade 3 ELLs (all ethnicity, strongest for Whites) tended to perform poorly compared with non-ELLs on phrases-in-context items, and grade 10 White ELLs tended to perform worse than White non-ELLs on main ideas. In contrast, grade 10 Asian, Hispanic, and White ELLs tended to outperform non-ELLs on evaluation items, suggesting a role for ethnicity and language background in reading assessment.

The above survey of the DIF literature reveals that, although some manifest variables such as gender and ethnicity can affect DIF in reading tests, the main cause of DIF is usually speculative or unidentified. Latent class models of DIF have shown that DIF is often occasioned by more complex dynamics than what the commonly adhered-to unifactorial methods can detect (Zumbo et al., 2015). A recent development in DIF analysis capable of capturing this dynamism is recursively partitioning Rasch trees (Strobl et al., 2015), which are discussed in more detail below.

2.3. Recursive partitioning rasch trees

The recursively partitioning Rasch trees approach (or just Rasch trees) to DIF is a trade-off between DIF across manifest variables and latent class models of DIF, such as the mixture Rasch models. Rasch trees combine "recursive partitioning" techniques and Rasch measurement (Strobl, Malley, & Tutz, 2009) and can identify groups of test takers displaying DIF. The test takers exhibiting DIF are specified by combining various manifest variables (covariates) such as demographic variables or language skills (Strobl et al., 2015). The tree-based part of the Rasch tree method is adopted from econometrics; data are partitioned recursively based on changes in data structure due to differences in the parameters between groups of test takers determined by combinations of manifest variables. The fact that the data are partitioned into several nodes, with items possessing different parameters in each node, indicates that one composite Rasch model cannot capture the complexity of the data (Strobl et al., 2015).

In Rasch trees, a Rasch model is initially fit to the entire sample, and then the stability of the test item difficulty parameters is evaluated across the pre-specified manifest variables. Where DIF is present, a significant lack of stability due to one or more manifest variables emerges; the manifest variable causing DIF is the most unstable of all variables, and factoring in its instability will lead to the most significant improvement of the fit of the model. Instability is defined as deviation of person parameters from the "overall mean" or zero, as estimated by the Rasch model. Generalized M-fluctuation tests (see Zeileis & Hornik, 2007) are used to evaluate the statistical significance of the deviation of the parameters from the overall mean. This process is recursive and generates multiple subsample stops only when splitting the subsamples results in no improvement to the fit of the model (Strobl et al., 2009). A test statistic and associated *p* value (with Bonferroni correction) are computed for each manifest variable; test statistics for categorical (including nominal) and continuous manifest variables are estimated using maximum Lagrange-multiplier statistic and the extension of the Lagrange-multiplier, respectively (Zeileis, Hothorn, & Hornik, 2008). In addition to the *p* value, which is used as one of the stopping rules in splitting, the minimum size of the sample in each node is considered as splitting proceeds. According to Strobl et al. (2009), the minimum size can be determined by the researcher depending on the qualities of the sample.

After selecting the suitable manifest variable(s) for splitting and generating the nodes, the model then establishes "cutpoints" where splits takes place. All possible cutpoints are tested to determine the optimal value where the sample can be branched (Strobl et al., 2015). Unlike conventional DIF where samples are divided into focal and reference groups, the splitting in Rasch trees requires no pre-determined manifest variables and the splits are obtained according to the data, rather than any pre-imposed manifest variable.

The recursive Rasch trees can be characterized as an advancing technique particularly well-suited for the third generation of DIF analysis (Zumbo, 2007; Zumbo et al., 2015). According to Zumbo, the third generation of DIF analysis, which has not yet attracted the research attention that it deserves in language assessment, is quite useful in investigating the cognitive processes of test takers in performing certain tasks and specifically in inspecting "lack of invariance". Rupp's (2005) approach to DIF is in line with the third generation and highly similar to the Rasch tree method. Rupp (2005, p. 91) argues that "parameter invariance is not a general property that holds exactly over all possible sets of subpopulations, however defined, as seems to be a common misperception among practitioners that use IRT [item response theory] models." In his analysis, Rupp first performed a CART analysis followed by an IRT analysis of each of the subpopulations which emerged in the CART analysis, an approach highly similar to the Rasch tree methodology. In the latter analysis, however, both estimates are performed within one framework and most of the attention is paid to the variables that induce DIF rather than the item-based magnitude of DIF. This, according to Zumbo (2007), assimilates such partitioning techniques with ANCOVA than DIF analysis, meaning that the detected discrepancies may not be necessarily treated as DIF but as a source of variance related to the construct—depending on the identified sources.

3. Method and data analysis

3.1. Data source and assessment

Item-level test data (with no missing cases) from 1550 university applicants (male = 899 or 58%; female = 651 or 42%) to a major university in Iran were used in the present study. The participants held master's degree in various fields such as engineering and social sciences and took the test as a part of the entrance examination for the PhD programs offered by the university. English is taught and learned as a foreign language in Iran and the participants should, therefore, be considered English as a foreign language (EFL) leaners.

The battery test administered to the applicants comprises several major sections, all in multiple-choice question format, modeled on the paper-based Test of English as a Foreign Language (TOEFL PBT). The sections are, as follows:

1. Grammar test, which assesses error identification ability and judgement of the most accurate grammatical form. It comprises two main forms: (A) 15 sentences with four underlined words, one of which is grammatically ill-formed (15 items), and (B) 10 sentences each containing a blank, which learners should fill in by choosing the most suitable of the four available options (10 items).
2. Vocabulary test, which assesses learners' knowledge of synonymy, polysemy, and collocations (major dimensions of depth of vocabulary). It comprises two main forms: (A) 15 sentences where learners should choose the most appropriate synonyms for the underlined words (15 items); and (B) 15 sentences where learners should choose the best words completing a collocation or indicating polysemy (15 items).
3. Reading comprehension test, which consists of six passages of different lengths (95–360 words) and aims to measure applicants' ability to determine viewpoints, understand explicit and implicit information, and draw conclusions and inferences (35 multiple-choice format items):
   a Passage 1: 243 words, 8 items
   b Passage 2: 96 words, 4 items
   c Passage 3: 171 words, 4 items
   d Passage 4: 177 words, 5 items
   e Passage 5: 144 words, 3 items
   f Passage 6: 359 words, 7 items
   g Sentence paraphrasing: 338 words, 4 items

3.2. Data analysis

3.2.1. Preliminary analysis of the psychometric quality

The WINSTEPS computer package, Version 3.80 (Linacre, 2014a)—Rasch measurement software—was employed to examine the unidimensionality and local independence of the grammar and vocabulary tests, alongside the fit of the items and the test takers to the Rasch model. This preliminary analysis was performed to identify potential sources of construct-irrelevant variance and aberrations in the data. Unidimensionality in each test was examined by submitting the Rasch model residuals (the differences between observed and expected values) to principal component analysis (PCAR) to identify any structures in the residuals, which, if substantive and significant, might jeopardize unidimensionality of the test. Local independence was also examined by correlating the linearized Rasch model residuals. Significantly high correlations between item residuals indicate certain degrees of local dependence (Linacre, 2014b). Infit and outfit mean square (MNSQ) indices were estimated per test—which are inlier and outlier sensitive chi-square values that identify anomalies closer to or farther away from item difficulty or person ability, respectively. The ideal range for these statistics is between 0.6 and 1.4 (Bond & Fox, 2015).

This round of Rasch measurement yielded lexical and grammatical ability measures per test taker, which were subsequently used as covariates in the Rasch trees analysis.

3.2.2. Rasch trees model

The Rasch trees analysis (Strobl et al., 2015) was applied to detect DIF in the reading comprehension test across gender, as well as with respect to potential differences in test takers' grammar and vocabulary abilities. This analysis was performed using the add-on package *psychotree* (Zeileis, Strobl, Wickelmaier, & Kopf, 2010), which is run using the R package (R Development Core Team, 2010). Where the analysis yields more than one node, the logistic Rasch model does not fit the data anymore; rather, there will be different item difficulty parameters per subgroup of test takers, as determined by their group memberships or the covariate(s). The tree structure and the nodes are also graphically presented by *psychotree*, facilitating the interpretation of the analysis.

Unlike conventional DIF analysis with manifest variables, the Rasch trees analysis is not carried out based on pre-specified variables. Instead, the algorithm learns to split the data into subsamples empirically through four primary steps: (1) estimation of item difficulty parameters for the entire sample using conditional maximum likelihood; (2) evaluation of parameter stability across the covariates; (3) using a Lagrange multiplier (LM) to divide the sample along the covariates where there is instability in the parameters, so as to maximize the fit of the model; and (4) repeating the three steps iteratively on the rest of the data to generate more subsamples along the covariates and eliminate instabilities (Strobl et al., 2009, 2015). The algorithm stops splitting the sample under two conditions: first, when no further instability is detected across item difficulty parameters for the levels of covariates—this is estimated using p value ($<$ .05)—and second, when the analyst sets a minimum subsample (node) size, at which point the algorithm is commanded not to split anymore.

Since Rasch trees constitute a recursive partitioning approach with high resemblance to the classification and regression trees (CART), this method might appear to need a pruning mechanism where some of the low branches of the tree are eliminated, as they do not fit the validation subsample; however, as Strobl et al. (2009) argue, the algorithm used for the Rasch trees analysis uses inferential statistics and p values, rather than the descriptive statistics commonly used in CART, thereby rendering it resilient to overfitting, which can affect the analysis in CART modeling. For the present study, item parameters across different covariates and their corresponding p values, the splitting points at each node, and the correlations between the nodes were estimated.

4. Results

4.1. Preliminary psychometric analysis

The grammar and vocabulary tests were submitted to Rasch modeling to investigate their reliability and psychometric features. Table 1 gives the infit and outfit MNSQ indices of the items on the tests, alongside their reliability and the number of eigenvalues in the PCAR. The fit of the items falls between 0.5 and 1.5 (except a negligible deviation in the outfit (1.51)), and item reliability and separation statistics are at the maximum. Person reliability, however, is approximately 0.7 in both tests, indicating that test takers have a fairly restricted range of lexico-grammatical abilities. This is not necessarily a problem and most likely reflects the features of the sample taking the test (Linacre, 2014b). Finally, the PCAR results provide evidence supporting the unidimensionality of the tests as the contrasts (dimensions) identified in the linearized residuals are below 2, suggesting that the observed structure in the residuals is not substantive enough to create an extra

Table 1
Psychometric Quality of the Grammar and Vocabulary Tests.

Test	Infit MNSQ min – max	Outfit MNSQ min – max	Item reliability (separation)	Person reliability (separation)	Eigenvalues in the 1st contrast (PCAR)
Grammar	.88–1.15	.75–1.26	1.00 (16.47)	.68 (1.45)	1.5
Vocabulary	.84–1.19	.70–1.51	1.00 (16.01)	.72 (1.60)	1.8

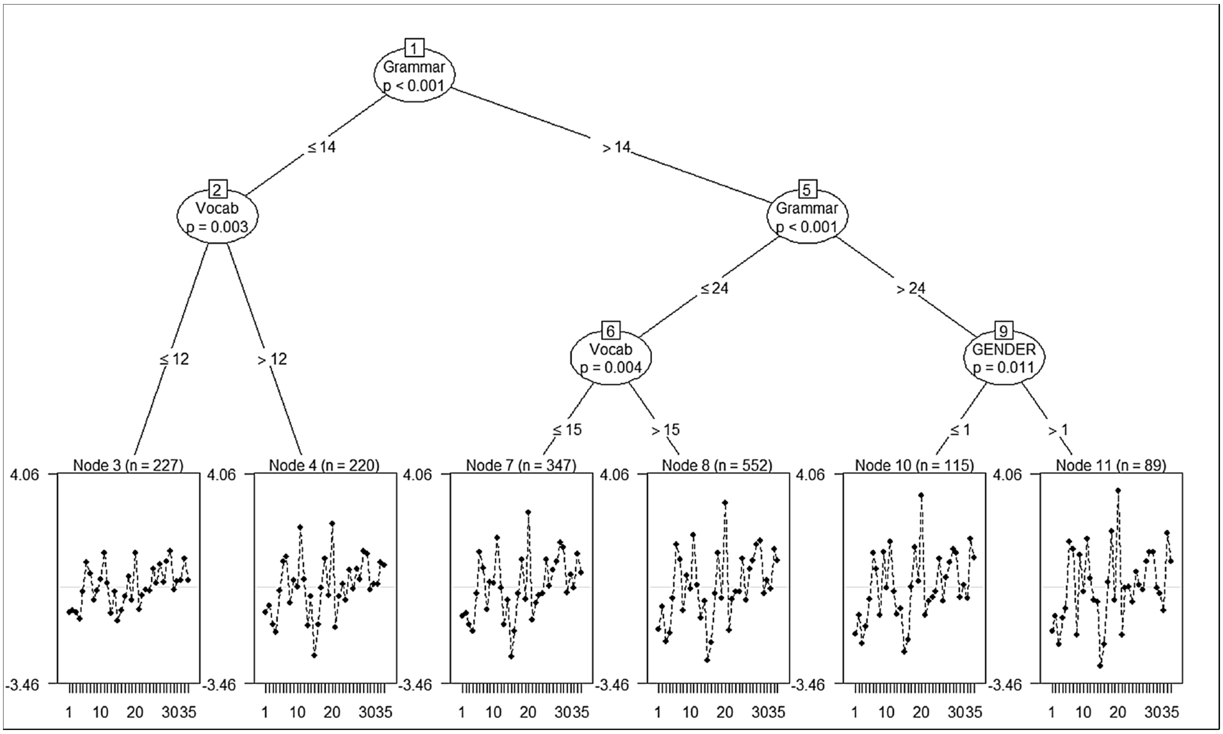

Fig. 1. Rasch tree for the reading test, with three covariates: vocabulary knowledge, grammar knowledge, and gender.

dimension (Linacre, 2014b).

4.2. Rasch trees results

The item-level data were evaluated using Rasch trees with respect to any possible group-specific discrepancies caused by the readers' vocabulary knowledge (measured by the vocabulary test), grammar knowledge (measured by the grammar test), gender, and academic major. The resulting Rasch tree model is presented in Fig. 1. The model partitions reading items based on the readers' grammar and vocabulary knowledge as well as gender, but not academic major, indicating that academic major does not induce any DIF in the data. There are 11 nodes, six of which are terminal nodes as represented by the rectangles at the bottom of the tree (nodes 3, 4, 7, 8, 10, and 11), where item difficulty parameters for the 35 reading items are graphically represented. In the top node, variation in grammar knowledge results in different item difficulty parameters for readers whose grammar test scores are ≤14 versus > 14 ($p < .001$); in other words, item difficulty parameters differ between readers with grammar scores > 14 and those with grammar scores ≤14. Next, readers with grammar test scores ≤14 are further partitioned based on their vocabulary knowledge into 227 readers with vocabulary test scores ≤12 in node 3 versus 220 readers with vocabulary test scores > 12 in node 4 ($p < .003$). Similarly, readers with higher grammatical scores in node 5 are further partitioned based on their vocabulary and gender in nodes 6 and 9, respectively. The partitioning leads to different item parameter estimations. For example, item 5 is easier for readers in nodes 3 and 4 than those in node 11. Gender affected only 204 (13.16%) test takers (i.e., 115 in node 10 and 89 in node 11), who had significantly high grammar scores. Overall, the observed partitioning patterns in item difficulty show that measurement invariance does not hold in the data and that the identified subgroups of readers are affected differently by DIF.

Table 2 presents the parameter instability evaluations with associated Bonferroni-adjusted p values for the four covariates across the 11 nodes. Overall, there are five nodes where the data are further split (nodes 1, 2, 5, 6, and 9), but, in the remainder of the nodes, splitting is halted since none of the covariates result in stable and statistically meaningful partitioning of the sample ($p > .05$). For example, in node 1, all covariates have significant p values (< .05), but the variable with the smallest value (grammar) is chosen for partitioning. The first splitting was based on the rule given in the right-hand column (≤14 and > 14) and culminated in two daughter nodes (2 & 5), where partitioning continued until the algorithm detected no further suitable covariates for partitioning the data (e.g., nodes 3 and 4). The cut-point for partitioning the sample based on grammar scores was not pre-specified and was determined by the algorithm. (Once the right rule for partitioning the data is found, the item difficulty parameters are then estimated across the resulting subsamples.)

In Table 3, the lowest item difficulty indices are underlined and the highest indices are emboldened. Most of the lowest and highest indices fall on nodes 3 and 11, which are the first and last external nodes. However, there is no distinctive pattern to differentiate these two nodes, i.e., whereas some of the extreme indices are the lowest in these nodes, others are the highest. The other columns in Table 3 give the mean scores and standard deviations (SDs) of the item difficulty indices across the nodes and the results of the *t*-tests estimating the statistical significance of the observed differences. Items 15 and 20 have the lowest (M = −2.325) and highest (M = 2.667) mean values, respectively, and items 10 (SD = 0.016) and 9 (SD = 0.727) have the lowest and highest SDs, respectively. This indicates that the dispersion of item difficulty is significantly high in item 9 and significantly low in item 10. The *t*-test results show that, except for items 10, 12, and 33, the other items (32 items) have significant fluctuations in item difficulty across the nodes. For a better representation, the scatterplot in Fig. 2 visualizes the fluctuations in item difficulty across the nodes.

5. Discussion

The present study applied recursive partitioning Rasch trees to a large-scale reading comprehension test to identify sources of DIF. The variables used to examine DIF were vocabulary knowledge, grammar knowledge, gender, and academic major. It was found that grammar and vocabulary induced discrepancies among test takers in a non-linear and recursive fashion. The readers were initially split into groups with

Table 2
Parameter Instability Tests: Test Statistics and their Corresponding p Values.

	Gender		Major		Grammar		Vocabulary		Rules for selected split (# of observations)
	Statistic	*p* value	Statistic	*p* value	Statistic	*p* value	Statistic	*p* value	
Node 1	74.901	.010	73.900	.014	**1.988680e + 02***	8.517133*e*-24	2.036758*e* + 02	6.574469*e*-23	≤14 | > 14 (1550)
Node 2	57.343	.477	46.812	.978	60.533	.290	**79.848***	.002	≤12 | > 12 (447)
Node 3	57.544	.463	47.36	.971	37.970	.999	46.06	.986	NA (227)
Node 4	45.65	.989	40.12	.999	52.32	.800	47.98	.960	NA (220)
Node 5	65.83	.101	66.48	.087	9.402124*e* + 01	**3.438628e-05***	9.184188*e* + 01	6.966001*e*-05	≤24 | > 24 (1103)
Node 6	57.934	.439	66.457	.088	55.186	.622	**78.466***	.004	≤15 | > 15 (899)
Node 7	51.058	.862	56.415	.538	45.623	.989	54.337	.677	NA (347)
Node 8	45.119	.992	64.485	.135	51.149	.857	50.304	.893	NA (552)
Node 9	**74.674***	.01	54.855	.639	47.699	.964	40.996	.999	1 | 0 (204)
Node 10	46.800	.977	50.884	.866	39.438	.999	44.85	.993	NA (115)
Node 11	58.192	.407	48.844	.932	41.666	.999	40.439	.999	NA (89)

*Note.** covariates selected for splitting the nodes are in bold print and marked by *.

low and high grammar knowledge (node 1); next, the low grammar readers were further split into readers with low and high vocabulary knowledge (node 2); readers in node 2 were then split into high vocabulary (node 3) and low vocabulary groups (node 4); on the other hand, high grammar readers were partitioned into low (node 6) and high ability, and the latter group was split again based on gender (node 9). This means that item difficulty parameters differed significantly for readers at each node, depending on the splitting variable (Rupp, 2005).

As Zumbo (2007) has argued, the third generation DIF analysis may not necessarily represent unfairness or bias in the test; rather, the observed discrepancies could point to a natural difference in the population which can be captured by certain covariates. As such, the results of the present study support previous reading research where grammar and vocabulary were found to be the main predictors of reading ability (Grabe, 2009; Koda 2005; Perfetti et al., 2005). Grammar and vocabulary knowledge are specifically important in bottom-up reading (Kintsch, 1998), and variation in these knowledge bases regulated readers' performance and likely caused differential bottom-up reading in low grammar and low vocabulary readers. Although the latter inference is rather speculative, in light of previous research, it seems plausible to make this assumption since low-ability readers tend to stick to the local meaning and parsing of texts, whereas

Table 3
Item Difficulty Parameters across the Nodes and the Associated p and t Values.

	node 1	node 2	node 3	node 4	node 5	node 6	node 7	node 8	node 9	node 10	node 11	Mean	*SD*	*t* value	*p* value
Item1	−1.215	−0.915	−0.911	*-0.905*	−1.364	−1.324	−1.062	−1.534	−1.650	**−1.685**	−1.591	−1.287	0.552	−13.978	.000
Item2	−0.828	−0.761	−0.837	*-0.666*	−0.839	−0.806	−0.949	−0.699	**−1.042**	−1.034	−1.039	−0.864	0.260	−21.262	.000
Item3	−1.502	−1.124	**−0.929**	−1.346	−1.720	−1.677	−1.352	−1.973	−2.049	−2.030	*-2.067*	−1.615	0.484	−13.505	.000
Item4	−1.527	−1.367	−1.155	−1.627	−1.598	−1.644	−1.608	*-1.661*	−1.293	−1.408	**−1.135**	−1.457	0.155	−24.610	.000
Item5	−0.334	−0.166	−0.168	**−0.151**	−0.387	−0.348	−0.235	−0.413	−0.596	−0.459	*-0.781*	−0.367	0.218	−6.243	.000
Item6	1.245	0.881	*0.859*	0.911	1.389	1.398	1.235	1.498	1.374	1.205	**1.617**	1.237	0.208	16.008	.000
Item7	0.822	0.759	*0.454*	1.073	0.864	0.847	0.660	0.965	0.946	0.652	**1.349**	0.854	0.298	11.940	.000
Item8	−0.781	−0.526	*-0.471*	−0.569	−0.888	−0.828	−0.799	−0.839	−1.293	−1.034	**−1.731**	−0.887	0.329	−8.074	.000
Item9	0.323	0.030	*-0.149*	0.218	0.458	0.298	0.170	0.384	1.191	**1.246**	1.143	0.483	0.727	3.295	.008
Item10	0.002	0.111	**0.267**	−0.017	−0.021	−0.003	0.131	−0.075	−0.100	−0.040	*-0.165*	0.008	0.016	.225	.826
Item11	1.723	1.634	*1.198*	**2.120**	1.770	1.811	1.740	1.859	1.665	1.631	1.730	1.716	0.346	25.949	.000
Item12	0.054	0.203	0.136	0.279	0.015	0.017	−0.046	0.066	0.024	*-0.173*	**0.284**	0.078	0.320	1.871	.091
Item13	−1.175	−1.155	−0.948	*-1.394*	−1.162	−1.231	−1.368	−1.123	−0.755	−0.968	**−0.486**	−1.070	0.301	−13.268	.000
Item14	−0.459	−0.281	**−0.188**	−0.359	−0.520	−0.489	−0.461	−0.497	−0.690	*-0.784*	−0.555	−0.480	0.301	−9.445	.000
Item15	−2.196	−1.695	**−1.232**	−2.471	−2.579	−2.586	−2.499	−2.652	−2.511	−2.322	*-2.837*	−2.325	0.105	−16.458	.000
Item16	−1.544	−1.084	**−0.837**	−1.370	−1.828	−1.807	−1.590	−1.991	−1.974	−1.905	*-2.067*	−1.636	0.378	−13.613	.000
Item17	−0.218	−0.214	*-0.359*	−0.056	−0.200	−0.250	−0.247	−0.241	0.072	0.004	**0.177**	−0.139	0.042	−2.811	.018
Item18	1.072	0.695	*0.382*	1.017	1.224	1.119	0.987	1.202	1.640	1.412	**1.970**	1.156	0.279	8.948	.000
Item19	−0.395	−0.394	−0.471	−0.303	−0.374	−0.428	−0.448	−0.405	−0.074	**0.214**	*-0.486*	−0.324	0.366	−5.092	.000
Item20	2.609	1.681	*1.198*	2.242	3.000	2.891	2.677	3.013	3.328	3.263	**3.439**	2.667	0.722	12.532	.000
Item21	−1.292	−1.104	**−0.801**	−1.469	−1.376	−1.387	−1.180	−1.547	−1.293	−1.034	*-1.731*	−1.292	0.308	−16.673	.000
Item22	−0.450	−0.337	−0.321	−0.341	−0.481	−0.509	*-0.576*	−0.455	−0.308	−0.510	**−0.047**	−0.394	0.120	−8.947	.000
Item23	−0.175	-0.010	−0.090	**0.080**	−0.224	−0.225	−0.297	−0.169	−0.202	*-0.360*	0.011	−0.151	0.311	−3.707	.004
Item24	−0.264	−0.309	−0.129	**−0.473**	−0.224	−0.201	−0.235	−0.169	−0.336	−0.173	*-0.555*	−0.279	0.212	−6.991	.000
Item25	0.835	0.608	0.632	0.598	0.935	0.977	0.970	0.991	0.793	**1.004**	*0.542*	0.808	0.287	14.755	.000
Item26	−0.210	0.010	**0.115**	−0.075	−0.282	−0.284	0.029	−0.480	−0.255	*-0.510*	0.067	−0.170	0.308	−2.623	.025
Item27	0.554	0.707	**0.800**	0.644	0.524	0.613	0.588	0.638	0.144	0.336	*-0.105*	0.495	0.218	6.133	.000
Item28	0.705	0.224	*0.179*	0.279	0.908	0.920	0.921	**0.930**	0.881	0.886	0.893	0.702	0.429	7.467	.000
Item29	1.362	1.105	*0.921*	1.282	1.465	1.528	**1.568**	1.520	1.282	1.328	1.245	1.328	0.033	22.565	.000
Item30	1.391	1.204	1.276	*1.159*	1.470	1.558	1.414	**1.645**	1.213	1.205	1.245	1.344	0.033	27.421	.000
Item31	−0.221	−0.118	−0.109	−0.113	−0.245	−0.245	−0.222	−0.249	−0.228	*-0.360*	**−0.047**	−0.196	0.175	−7.291	.000
Item32	0.195	0.131	0.201	0.080	0.240	0.307	*0.448*	0.237	−0.074	0.047	*-0.226*	0.144	0.023	2.590	.027
Item33	−0.078	0.141	**0.223**	0.080	−0.146	−0.065	−0.034	−0.075	−0.596	−0.409	*-0.863*	−0.166	0.346	−1.672	.126
Item34	1.261	0.924	1.019	*0.860*	1.394	1.290	1.179	1.360	1.795	1.723	**1.908**	1.338	0.610	12.697	.000
Item35	0.711	0.512	*0.245*	0.786	0.802	0.766	0.490	0.939	0.968	**1.044**	0.893	0.741	0.182	10.261	.000
Splitting Covariate	Grammar	Vocabulary	–	–	Grammar	Vocabulary	–	–	Gender	–	–	–	–	–	–

Note. The lowest difficulty indices are *underlined* and the highest are **emboldened**.

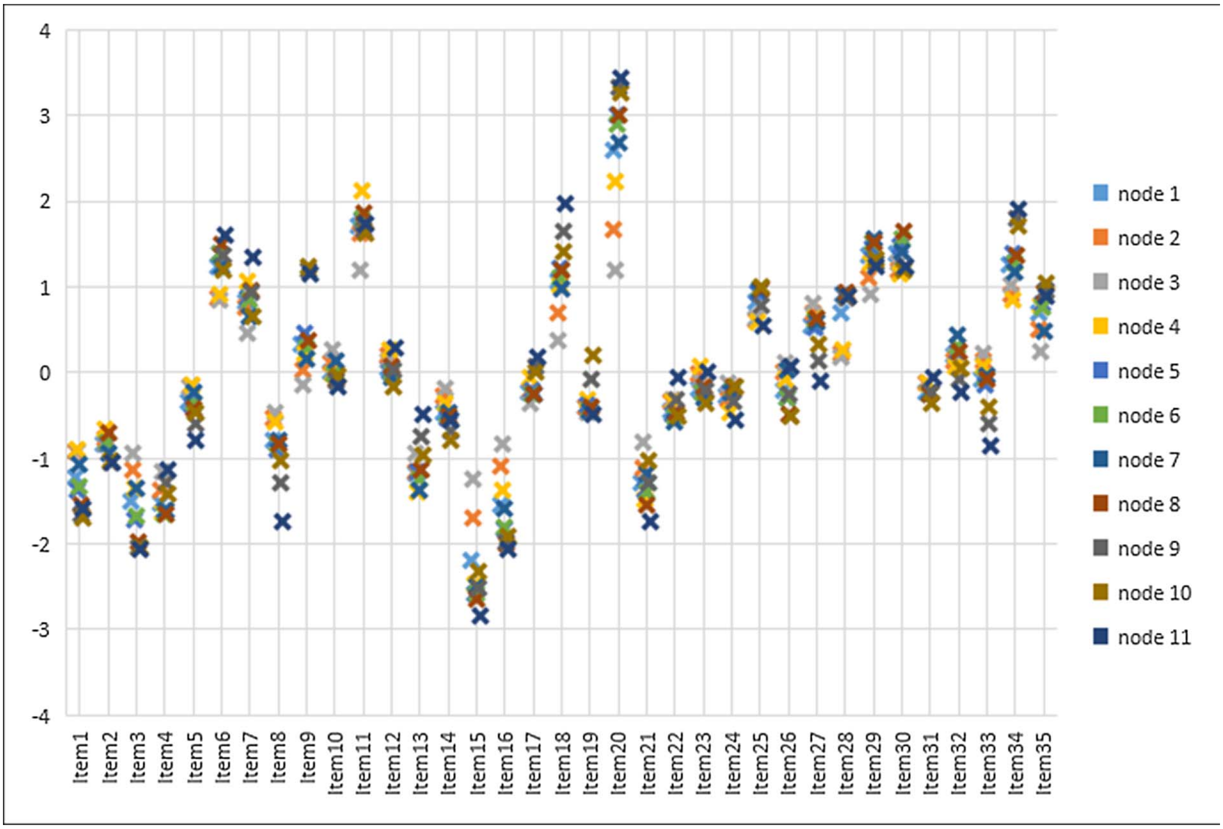

Fig. 2. Item difficulty parameters across the nodes.

high ability readers rely chiefly on top-down processing of the text (Aryadoust & Zhang, 2016). A question that should be addressed in future research is whether and how factoring in other mechanisms and variables—such as metacognitive variables and general knowledge (which were considered by Aryadoust and Zhang) —would reveal more information about the reading processes of the test takers.

With regards to the significance and implication of the DIF-causing variables, the findings showed that once data were partitioned by the Rasch trees algorithm, different permutations of variables emerged. As a general finding, groups with lower vocabulary and grammatical knowledge seem to be “disadvantaged”; however, since these variables are construct-relevant and sub-dimensions of reading, it is still fair to compare the ability level of the groups. Nevertheless, for the 13.16% of test takers affected by gender (nodes 10 and 11), it is important to carry out post-hoc analysis. It is suggested that these test takers be identified and DIF analysis using the logistic Rasch model be carried out to determine the items that caused DIF. This technique (using the logistic Rasch model to determine DIF across discrete manifest variables) can be used in future research as a useful supplement for Rasch trees.

Like previous DIF research with manifest variables (Gnaldi & Bacci, 2016), the present study found a significant role for gender. However, unlike these studies, it was found that gender-based DIF affected only a small portion of the sample. This is partly in line with the results of latent class DIF analysis research used by Aryadoust and Zhang (2016), Chen and Jiao (2014), and Hong and Min (2007) who found that gender had a weak or absent role in generating DIF in reading. It is suggested that the role of other manifest variables such as ethnicity, age, and language background in yielding DIF should be revisited via Rasch trees.

Turning to the methodological overtones of Rasch trees, it is important to note that DIF groups and the cutpoints in Rasch trees were formed by combining the manifest variables (e.g., vocabulary, grammar, and gender); however, unlike traditional DIF analysis, these variables are not pre-specified. According to Strobl et al. (2015, p. 293), “[t]his is a key feature of the model-based recursive partitioning approach employed here, which makes it very flexible for detecting groups with DIF and distinguishes it from parametric regression models, where only those main effects and interactions that are explicitly included in the specification of the model are considered.” The application of Rasch trees in other assessment contexts, such as listening, should also be investigated. One direction for future development is the extension of Rasch trees to performance data, where subjective performance of raters has been researched extensively. If certain variables can be used to partition raters and adjust their bias terms, the results might become more reliable and specific.

Another methodological note is that the PCAR in this study indicated that a unidimensional model fits the data, but the results of the subsequent Rasch trees analysis indicated DIF, which is frequently a threat to unidimensionality. These seemingly contradictory results are due to the insensitivity of PCAR to multidimensionality caused by DIF. PCAR is a suitable technique to identify construct-irrelevant factors based on the correlations of the Rasch model’s residuals, but as previous research has shown, it cannot identify the dimensions that induce DIF (see Aryadoust, 2012). The reason is that DIF is based on the differences between the probabilities of answering test items correctly rather than residuals. Rasch measurement is well-suited to investigate these two sources of construct-irrelevant variance.

6. Conclusion

This study set out to investigate DIF in reading using Rasch trees. Overall, the observed DIF caused by grammar and vocabulary ability was not treated as a sign of multidimensionality of the reading test; rather, it supported theories of reading that stress the significance of these variables, which is consistent with the third generation of DIF detection (Zumbo, 2007). Gender, nevertheless, caused multidimensionality in a small region of the covariates' space, which we take as evidence that only a relatively small subgroup was affected by a source of CIV.

The present study has implications for reading pedagogy and research. The first implication is that lexical and grammatical knowledge do play a significant part in second language reading, and it is therefore important to allocate some time in reading classes and courses to teaching vocabulary and grammar. Learners should be made aware of the importance of these variables, and they should be assisted in their efforts to improve their vocabulary and grammar. Lastly, Rasch trees represent a recent development in psychometric analysis and more models of this type should be developed to help researchers enhance the precision of their analyses. For example, Tutz and Burger (2015) developed an item-specific recursive partitioning tree method where item-specific information can be extracted. Another recent development is the structural equation modeling trees technique, which is used for cause-effect research (Brandmaier, Oertzen, McArdle, & Lindenberger, 2013). These innovations facilitate empirical exploration of data and significantly enhance the precision of the results.

References

Alderson, J. C. (2000). *Assessing reading.* Cambridge, UK: Cambridge University Press.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185–198.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. Guthrie (Ed.). *Comprehension and teaching: Research reviews* (pp. 77–117). Newark, DE: International Reading Association.

Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of IELTS listening test. *International Journal of Listening, 26*(1), 40–60.

Aryadoust, V., & Zhang, L. (2016). Fitting the mixed Rasch model to a reading comprehension test: Exploring individual difference profiles in L2 reading. *Language Testing, 33*(4), 529–553.

Bacci, S., Bartolucci, F., & Gnaldi, M. (2014). A class of multidimensional IRT models for ordinal polytomous item responses. *Communications in Statistics: Theory and Methods, 43*(4), 787–800.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika, 72*, 141–157.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). London: Lawrence Erlbaum.

Brandmaier, A. M., Oertzen, T. V., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods, 18*, 71–86.

Bügel, K., & Buunk, B. P. (1996). Sex differences in foreign language text comprehension: The role of interests and prior knowledge. *The Modern Language Journal, 80*, 15–31.

Cadime, I., Viana, F. L., & Ribeiro, I. (2014). Invariance on a reading comprehension test in European Portuguese: A differential item functioning analysis between students from rural and urban areas. *European Journal of Developmental Psychology, 11*(6), 754–766.

Chen, Y.-F., & Jiao, H. (2014). Exploring the utility of background and cognitive variables in explaining latent differential item functioning: An example of the PISA 2009 reading assessment. *Educational Assessment, 19*, 77–96.

Chiu, M. M., & McBride-Chang, C. (2006). Gender, context, and reading: A comparison of students in 43 countries. *Scientific Studies of Reading, 10*, 331–362.

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

Florit, E., & Cain, K. (2011). The simple view of reading: Is it valid for different types of alphabetic orthographies? *Educational Psychology Review, 23*, 553–576.

Gnaldi, M., & Bacci, S. (2016). Joint assessment of the latent trait dimensionality and observed differential item functioning of students' national tests. *Quality & Quantity, 50*(4), 1429–1447.

Golden, R. M., & Rumelhart, D. E. (1993). A parallel distributed processing model of story comprehension and recall. *Discourse Processes, 16*, 203–237.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* Cambridge: Cambridge University Press.

Hong, S., & Min, S.-Y. (2007). Mixed Rasch modeling of the self-rating depression scale: Incorporating latent class and Rasch rating scale models. *Educational and Psychological Measurement, 67*, 280–299.

Hu, M. H.-C., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*, 403–430.

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning, 64*, 160–212.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, UK: Cambridge University Press.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach.* New York: Cambridge University Press.

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, 31*(1), 89–109.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren, & M. Nordman (Eds.). *Special language: From humans thinking to thinking machines* (pp. 316–332). Clevedon, England: Multilingual Matters.

Linacre, J. M. (2014a). *A user's guide to WINSTEPS.* Chicago, IL: Winsteps.com.

Linacre, J. M. (2014b). *WINSTEPS, version 3.80 [Computer program].* Chicago, IL: Winsteps.com.

McGeown, S., Goodwin, H., Henderson, N., & Wright, P. (2012). Gender differences in reading motivation: Does sex or gender identity provide a better account? *Journal of Research in Reading, 35*, 328–336.

Ownby, R. L., & Waldrop-Valverde, D. (2013). Differential item functioning related to age in the reading subtest of the test of functional health literacy in adults. *Journal of Aging Research, 2013*.

Pae, H. (2011). Differential item functioning and unidimensionality in the Pearson Test of English Academic. Pearson Education Ltd. Retrieved from https://pearsonpte.com/wp-content/uploads/2014/07/RN_Differential-ItemFunctioning.pdf.

Paris, S. G., & Winograd, P. (1990). How metacognition can promote academic learning and instruction. In B. F. Jones, & L. Idol (Eds.). *Dimensions of thinking and cognitive instruction* (pp. 15–51). Hillsdale, NJ: Lawrence Erlbaum.

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*, 357–383.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skills. In M. J. Snowling, & C. Hulme (Eds.). *The science of reading: A handbook* (pp. 227–247). London, England: Blackwell.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*, 22–37.

Purpura, J. (2004). *Assessing grammar.* Cambridge: Cambridge University Press.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*, 513–536.

R Development Core Team (2010). *R: A language and environment for statistical computing.* [Retrieved from http://www.R-project.org/].

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271–282.

Rupp, A. A. (2005). Quantifying subpopulation differences for a lack of invariance using complex examinee profiles: An exploratory multigroup approach using functional data analysis. *Educational Research and Evaluation, 11*(1), 71–97.

Stanovich, K. E., & Stanovich, P. J. (2006). Fostering the scientific study of reading instruction by example. In K. Dougherty Stahl, & M. McKenna (Eds.). *Reading research at work: Foundations of effective practice* (pp. 36–44). New York, NY: Guilford Press.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika, 80*(2), 289–316.

Strobl, C., Malley, J., & &Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods, 14*, 323–348.

Tutz, G., & Berger, M. (2015). Item focused trees for the identification of items in differential item functioning. *Psychometrika*. http://dx.doi.org/10.1007/s11336-015-9488-3 [published online].

von Davier, M., & Rost, J. (1995). Polytomous mixed rasch models. In G. H. Fischer, & I. W. Molennar (Eds.). *Rasch models: Foundations, recent developments, and applications* (pp. 371–379). New York: Springer Verlag.

Yano, Y., Long, M. H., & Ross, S. (1994). The effects of simplified and elaborated texts on foreign language reading. *Language Learning, 44*, 189–219.

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., Strobl, C., Wickelmaier, F., & Kopf, J. (2010). *Psychotree: Recursive partitioning based on psychometric models. R package version 0. 11-1.* [URL http://CRAN.R-project.org/package=psychotree].

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223–233.

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Astivia, O. L. O., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12*(1), 136–151.