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

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-pre-specified 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 pre-specified 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.

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