USING RASCH TREE TO DETECT UNIFORM DIFFERENTIAL ITEM FUNCTIONING AND ITEM DIFFICULTY PARAMETERS IN THE PROGRESS IN INTERNATIONAL READING LITERACY DATA (PIRLS)

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE DOCTOR OF PHILOSOPHY

BY

HEATHER JEFFERS

ADVISOR: DR. HOLMES FINCH

BALL STATE UNIVERSITY

MUNCIE, IN

May 2020
# TABLE OF CONTENTS

Abstract......................................................................................................................................................iv

Acknowledgements........................................................................................................................................v

List of Tables..................................................................................................................................................vi

List of Figures................................................................................................................................................vii

Chapter 1: Introduction...................................................................................................................................1

Chapter 2: Literature Review..........................................................................................................................7
  Reading and Gender Differences..................................................................................................................7
  PIRLS and Reading Literacy.........................................................................................................................8
  Classical Test Theory..................................................................................................................................10
  Item Response Theory................................................................................................................................11
    Item Characteristic Curve..........................................................................................................................14
    Assumptions of IRT..................................................................................................................................19
  Differential Item Functioning (DIF)...........................................................................................................20
    Types of DIF............................................................................................................................................23
    Methods for assessing DIF.......................................................................................................................25
      Mantel-Haenszel..................................................................................................................................25
      Logistic Regression..............................................................................................................................26
      Item Response Theory Likelihood Ratio..............................................................................................26
      SIBTEST..............................................................................................................................................28
      RaschTree.........................................................................................................................................29
    Comparison of Methods.........................................................................................................................33
  Research Questions......................................................................................................................................36
Chapter 3: Method .......................................................................................................................... 37
Chapter 4: Results .......................................................................................................................... 39
    RaschTree for Gender .................................................................................................................. 40
    RaschTree for Reading Engagement .......................................................................................... 42
    RaschTree for All Variables ...................................................................................................... 44
Chapter 5: Discussion .................................................................................................................... 53
References .................................................................................................................................... 60
ABSTRACT

This study presents an alternative method, Rasch Tree, for detecting differential item functioning and item difficulty parameters with an existing dataset, the Progress in International Reading Literacy Data (PIRLS). At present, there are methods of DIF detection, but each method has shortcomings and struggles depending on the dataset.

Rasch Tree is presented as a method which may be able to detect both DIF and difficulty parameters in a dataset with multiple variables, both dichotomous and ordinal, as well as interactions between all variables included, without having to fit multiple models or pre-set cut-points. Rasch Tree was able to not only detect DIF and item difficulty for a dichotomous variable, but was also able to detect interactions between a dichotomous variable and multiple ordinal variables.
ACKNOWLEDGEMENTS

Thank you first and foremost to my chair, Dr. Holmes Finch. Not only for the words of wisdom, advice, and help with every step of this program, but also for your patience. I started this program unsure of my own ability to finish it and if I had a chair who was any less supportive, I have no doubt I would not have been able to finish it. I will always be thankful for that and will always try to be a good representation of what I’ve learned from you as a researcher, as a teacher, and as a person. Also, thank you to my committee - Drs. Tagler, Fletcher, and Bolin - I know this was a long process, but I am so thankful to you all for still pushing me and working with me through the years.

To my parents, Curt and Sherry Jeffers, thank you for always pushing and believing in me. If it weren’t for you both, I never would have pursued something that seemed like such a wild stretch. I love you both.

Tyler, your patience and love through this very long process kept me going (especially these last few months) through early mornings, late nights, and every last push to finish this up. You did everything you could to make this as easy as possible on me and that is something that did not go unnoticed. Thank you for motivating me to finish. Thank you for talking me down when I got overwhelmed. Thank you for taking other things off my plate so I could work on this. Thank you for everything.

Finally, to my downtown Muncie family, thank you for... well, everything. From the late nights going over papers with me, having food delivered so I didn’t leave and get distracted, breaks when they were much needed, pushing me when I wanted to give up, and everything else in between. I could not have done this without every single one of you. Your support means the world to me.
LIST OF TABLES

Table 1. Percent correct for PIRLS reading comprehension items.................................41
Table 2. Means and standard deviations for attitude towards reading scales.....................42
Table 3. Item difficulty parameters.................................................................................44
Table 4. Item difficulty parameters of ordinal variables..................................................47
Table 5. Combined variables item difficulties..................................................................51
LIST OF FIGURES

Figure 1. Item Characteristic Curve for Rasch Model.................................15
Figure 2. Item Characteristic Curve for 1PL..............................................16
Figure 3. Item Characteristic Curve for 2PL..............................................17
Figure 4. Item Characteristic Curve for 3PL..............................................18
Figure 5. Item Characteristic Curve for Uniform DIF.................................23
Figure 6. Item Characteristic Curve for Non-Uniform DIF...........................24
Figure 7. Example of RaschTree...............................................................31
Figure 8. RaschTree for Gender...............................................................43
Figure 9. RaschTree for Ordinal Variables (Attitude Scale).........................46
Figure 10. RaschTree for all Variables (Gender, Attitude Scale, and Combined)........50
CHAPTER 1

INTRODUCTION

Differential Item Functioning

There is rising concern in fairness in educational and psychological assessment (Karami & Salmani Nodoushan, 2011). Differential Item Functioning (DIF) has been explored as one possible explanation of what may cause bias in an assessment and in fact has been described as the ‘standard in psychometric bias analysis’ (Zumbo, 1999). Differential item functioning (DIF) detection is used in educational and psychological testing to determine the probability of an item being answered incorrectly after controlling for ability level (Roussos & Stout, 1996). Simply put, DIF is present if two students of the same ability level of the trait being measured (for example, reading comprehension) have different likelihoods of answering a test question correctly.

Because very important decisions regarding academic placement, education plans, and even resource allotment can be decided based on scores to these assessments, it is important to understand not only what exactly is being measured, but also any issues that arise from the scale development and validation. Any assessment that has any items with significant DIF may be unfair for certain groups. With the impact of these scales being as large as they are, it is important that the items that are showing DIF be identified. Once they are identified the item can then be improved by test developers or deleted from the assessment (Westers & Kelderman, 1992).

Currently, a variety of statistical tests are used to detect DIF. Some of these methods compare a focal and reference group, such as males and females. These latent class approaches, namely the Mantel-Haenszel (MH; Mantel & Haenszel, 1959), the logistic
regression (LR) approach (Swaminathan & Rogers, 1990), the Thissen, Steinberg, and Wainer (1988) DIF detection strategy (IRTLR, the Item Response Theory Likelihood-Ratio Test (IRTLRT), SIBTEST (Shealy & Stout, 1993), & Rasch trees (Strobl, Kopf, & Zeileis, 2013), allow DIF detection in previously unknown groups. Current literature about these DIF detection strategies does not give overwhelming evidence for one specific test and is dependent on the conditions simulated in the studies. As of now, the detection strategies previously mentioned have been negatively affected by an increased Type I error rate by sample size, model being tested, parameters included in the model, percentage of DIF present, length of the assessment, differences in abilities of test takers, and interaction between any combination of these variables (Atalay Kabasakal, Arsan, Gok, Kelecioglu, 2014).

Typically, there are variables that are proposed to be tested for DIF, including age, gender, or ethnicity (example citations). The issue with this method of predetermined cut-points, however, is that if later analyses find a difference between groups that was not tested, it is possible that this DIF detection could just be due to previously unnoticed DIF (Strobl, Kopf, & Zeileis, 2015). Another issue with most approaches to DIF detection is that there is a loss of information due to variables that are continuous needing to be changed into different types of variables.

RaschTree

A new semi-parametric approach for DIF detection in Rasch Models has been recently proposed by Stobl, Kopf, and Zeileis (2013) and is based in recursive partitioning. Using the new model, covariates are used to detect model parameter differences. The newly proposed Rasch method is based in newer statistical methods and therefore can be
more robust than classification and regression trees (CART), which is the current approach that is most used to detect DIF. Using this new method, results of DIF detection are presented visually, which is a main advantage over preexisting DIF detection strategies. Variables also do not need to be specified into an algorithm, which makes it more exploratory and therefore flexible than DIF detection strategies that are parametric. A cut-point between the focal and reference group does not have to be pre-specified, rather, an affiliated cut-point with the largest parameter difference is detected automatically. Because of this, DIF will be detected for covariates. Also, as a result of the produced visual in the form of a Rasch Tree with nodes that split to demonstrate the cut-points, dimensions affected by DIF are easily identified (Strobl, Kopf, & Zeileis, 2010).

The main issue with the previously mentioned DIF detection strategies that the RaschTree Method attempts to remedy is that predefined groups must be used. With the currently proposed strategy, non-predefined group differences can be detected and explained. Although latent class approaches can detect DIF in non-predefined groups, they are not directly interpretable. Raschtree also seeks to remedy this problem by allowing for detection of latent groups and non-predefined groups and remains interpretable. These advantages over other DIF analyses make it more useful than previous methods. Simulations over multiple conditions comparing DIF detection strategies have shown that a more efficient test for DIF that minimizes Type I error and increased power needs to be implemented, as individual tests have major limitations depending on parameters and conditions simulated.

The current literature about different DIF detections strategies is lacking in that there is no answer for non-specified cutpoints (exploratory approach) and there is no way
for most detection strategies to measure interactions between groups, therefore limiting exactly what source of DIF in a question may be. The present study seeks to remedy these issues by investigating whether RaschTree is a good way to detect DIF in non-specified focal and reference groups and take a more exploratory approach, as well as to try to detect DIF in interactions between groups and produce easily interpretable results.

**DIF Detection**

Classical test theory (CTT) and item response theory (IRT) are both often used in educational assessment and measurement. CTT focuses on an individual’s total score, including an error term and true score, while IRT focuses on individual item performance while including parameters that vary depending on the model used. Individual ability level, item difficulty, and item discrimination are included with parameter estimation differing depending on which model is being used.

Item response theory has certain assumptions that must be met, including unidimensionality, local independence, and monotonicity (deAyala, 2009). Unidimensionality states that the trait of interest, or the latent trait, is unidimensional and measures only one thing. In contrast to CTT, in IRT there is no variability in the item parameters. IRT is more practical to use because of this assumption of unidimensionality, as well as easier to use, given that once the discrimination and difficulty parameters are set based on the sample they do not need to be estimated. Although the assumption of unidimensionality is discussed, it is specific to only unidimensional models, and multidimensional models do exist and can be used (Shealy & Stout, 1993).

Local independence is the second assumption of IRT models and is directly related to the first assumption of unidimensionality. It states that after the latent trait is controlled
for, the item responses are not related to one another. A response to one specific item does not influence the individual’s response to another item on the assessment. If they are related and there is an influence or relationship between responses, more than one latent trait is being measured, which indicates that the assumption of unidimensionality is also violated. The third assumption of IRT is monotonicity and assumes that an increase in the latent trait being measured will have a positive impact on the likelihood of a correct response to an item.

Item characteristic curves are an important part of IRT based methods. They demonstrate the expected and observed values for each item on an assessment. If the assumption of monotonicity holds true, the graph will be linear. The steeper a line in an ICC, the better an item discriminates between high and low ability. Conversely, the flatter a line, the lower the discrimination value of the item.

Currently, a variety of statistical tests are used to detect DIF. Some of these methods compare a focal and reference group, such as males and females. These latent class approaches, namely the Mantel-Haenszel (MH; Mantel & Haenszel, 1959), the logistic regression (LR) approach (Swaminathan & Rogers, 1990), the Thissen, Steinberg, and Wainer (1988) DIF detection strategy (IRTLR, the Item Response Theory Likelihood-Ratio Test (IRTLRT), SIBTEST (Shealy & Stout, 1993), & Rasch trees (Strobl, Kopf, & Zeileis, 2013), allow DIF detection in previously unknown groups. Current literature about these DIF detection strategies does not give overwhelming evidence for one specific test and is dependent on the conditions simulated in the studies. As of now, the detection strategies previously mentioned have been negatively affected by an increased Type I error rate by sample size, model being tested, parameters included in the model, percentage of DIF
present, length of the assessment, differences in abilities of test takers, and interaction between any combination of these variables (Atalay Kabasakal, Arsan, Gok, Kelecioglu, 2014).

A new semi-parametric approach for DIF detection in Rasch Models has been recently proposed by Stobl, Kopf, and Zeileis (2013) and is based in recursive partitioning. Using the new model, covariates are used to detect model parameter differences. The newly proposed Rasch method is based in newer statistical methods and therefore can be more robust than classification and regression trees (CART), which is the current approach that is most used to detect DIF. Using this new method, results of DIF detection are presented visually, which is a main advantage over preexisting DIF detection strategies. Variables also do not need to be specified into an algorithm, which makes it more exploratory and therefore flexible than DIF detection strategies that are parametric. A cut-point between the focal and reference group does not have to be pre-specified, rather, an affiliated cut-point with the largest parameter difference is detected automatically. Because of this, DIF will be detected for covariates. Also, as a result of the produced visual in the form of a Rasch Tree with nodes that split to demonstrate the cut-points, dimensions affected by DIF are easily identified (Strobl, Kopf, & Zeileis, 2010).

The main issue with the previously mentioned DIF detection strategies that the RaschTree Method attempts to remedy is that predefined groups must be used. With the currently proposed strategy, non-predefined group differences can be detected and explained. Although latent class approaches can detect DIF in non-predefined groups, they are not directly interpretable. Raschtree also seeks to remedy this problem by allowing for detection of latent groups and non-predefined groups and remains interpretable. These
advantages over other DIF analyses make it more useful than previous methods. Simulations over multiple conditions comparing DIF detection strategies have shown that a more efficient test for DIF that minimizes Type I error and increased power needs to be implemented, as individual tests have major limitations depending on parameters and conditions simulated. The current method proposed is meant to fill the gap that other DIF detection strategies leave, including being able to include multiple variables, detect interactions between variables, and produce easily interpretable results.

CHAPTER 2
LITERATURE REVIEW
Reading and Gender Differences

In terms of general intelligence, there does seem to be no significant differences between genders (Halpern, 2000). However, there are usually observable differences in specific cognitive abilities, and in reading both at the national and international level there are consistently gender differences noted (Mullis et al., 2003; Mullis et al., 2007; Lynn & Mick, 2009; Reilly, 2012). Although the discrepancies in reading achievement have long been a research question of interest and until recently was most concerned with the disadvantages of girls, boys have (as of late) become the focus of attention for underachievement (Hochweber & Vieluf, 2018). Caplan and Caplan (2016) have proposed the idea that gender differences were never actually present or whether the differences were due to bias. If there are differences related to bias, this could be directly linked to DIF and assessment fairness. According to Reilly, Nuemann, & Andrews (2019), girls not only significantly outperformed boys in reading ability, but those differences tended to have a larger effect size as students got older. While the PIRLS did measure younger grades, this is
an important aspect to consider when looking at the results of the PIRLS data. If there is bias that is being found in middle school-aged children it is possible that the larger gap in older grades are also due to DIF.

PIRLS and Reading Literacy

The concept of ‘reading to do’ (Stiggins, 1982) or being able to use written text to attain a goal, is the framework in which the PIRLS is grounded. Rather than focusing on just a student’s ability to comprehend texts, it is now more than ever important that students can demonstrate that they are able to apply information gained from the texts (Coulombe, Tremblay, & Marchand, 2004). Based on an IEA (1991) study, reading literacy was defined as ‘the ability to understand and use those written language forms required by society and/or valued by the individual’, and that is the definition the PIRLS is based on. Today, that definition has evolved to include students’ experiences, the importance of reading in not only school but everyday life, and the variety of text available in today’s society versus in 1991. Currently, the PIRLS definition of reading literacy is:

Reading literacy is the ability to understand and use those written language forms required by society and/or valued by the individual. Readers can construct meaning from texts in a variety of forms. They read to learn, to participate in communities of reading in school and everyday life, and for enjoyment.

Because reading is an interactive process (Anderson & Pearson, 1984; Kintsch, 2013; Ruddell & Unreau, 2004), the adapted definition for PIRLS is important in that it includes that aspect of interaction along with how they may create meaning and also how involvement can affect reading literacy.
An important aspect of using texts to understand and construct meaning and use the information presented is the skills and cognitive strategies they have. This has been presented in all stages of reading: before, during, and after (Kintsch, 2013; Pressley & Gaskins, 2006; Miller & Faircloth, 2009, Baker & Beal, 2009). Context of reading can help support engagement and motivation to read but can also have negative effects that make it harder for students to construct meaning (Christianson & Luke, 2011; Miller & Faircloth, 2009). The attitude scale of the PIRLS did measure engagement and motivation, as well as enjoyment and attitude, to look at the effect those things may have on assessments.

A new factor to consider that is also now included in the definition of reading literacy for this particular project is the increase in text types. While at one point traditional texts such as books, other documents, and even digital forms such as email and websites were the norm, there are many different text types today that can be and are used in education (Leu, Kinzer, Coiro, Castek, & Henry, 2013; Reuda, 2013). For example, there may be a mixture of traditional text types combined newer forms, such as videos or other multimedia. This is something that must be taken into consideration because the way that one may process one form of multimedia may be better than another and therefore affect scores on assessment while having nothing to do with actual ability level.

Social interaction is another way that reading and constructing meaning has changed over time. While context before may have meant learning in the classroom by reading and reflecting on one's own, there are now more communities with which to engage, which can affect how one creates meaning. Not only may students be more apt to have conversations with peers about reading, but with the integration of multimedia and
different ways of reading outside of the classroom (e-learning and online forums), students are interacting with groups not only in a more hands-on way than before but also with groups who they may not normally interact with.

Knowing what we now know about the evolution of the definition of reading literacy and what it includes over time, we can see that the PIRLS is an important aspect of learning more about reading literacy not only internationally, but also over time given changes that may affect reading literacy.

**Classical Test Theory**

Educational measurement and assessment often employs classical test theory (CTT) and item response theory (IRT). Classical test theory measures an individual's score on an assessment as a function of two parts: the true score, which is the test taker's actual ability or knowledge, and error, which affects the score but does not reflect the actual ability of the individual. Classical test theory (CTT) is expressed: $X = T + E$.

In the CTT equation, the true score, $T$, is the expected number correct score over an infinite number of independent administrations of a specific test and would be obtained if there were no errors in measurement. Unfortunately, because there is always error present in testing characteristics and administration, $E$ must be added to an individual's true score to represent that. Because there is always error, those two terms together represent an observed score, $X$.

Given what exactly CTT is measuring, it is important to note that the quality of the test is also being measured. If there is an error value, it is important to understand why there was error and where it is coming from. If it is due to the assessment, that must be
adjusted and re-tested accordingly to try to get as close as possible to measuring only the trait of interest. However, if it is not due to the quality of the assessment, that is when we begin to look at if it may be due to DIF. However, because of the nature of CTT, Item Response Theory is often used in testing and assessment as a means to have information based on each individual item, which can paint a more whole picture of both the test and where error is coming from.

**Item Response Theory**

Item response theory, unlike CTT, does not focus on an overall score on an assessment, but rather each individual item performance. The IRT model estimates the likelihood that a person will choose the correct response to a specific item given their ability level and the difficulty and discrimination of the item.

IRT assesses:

\[ P(\text{correct}) = P(\theta, a, b, c) \]

This equation takes into account the individual’s ability level and item parameters a, b, and c (Embretson & Reise, 2000) when estimating the probability of the examinee getting an item \( X \) correct. In this case, difficulty, discrimination, and ability level are all taken into account.

The 1 parameter logistic model is similar to the Rasch model, with the main difference being that rather than the discrimination being set to 1 the person sample is parameterized by a mean and standard deviation for item estimation. This can be defined as:

\[ P(\text{correct}) = \frac{e^{(\theta - b)}}{1 + e^{(\theta - b)}} \]
where \( \theta \) represents ability and \( b \) represents item difficulty. It is exactly the same as the Rasch model with the exception that item discrimination is not set equal to 1, but is rather estimated from the data. A single discrimination parameter value is estimated for all of the items.

The Rasch Model is a special case of the 1-parameter logistic model, with discrimination set to 1. The objective of the Rasch Model is to obtain a measurement of latent variables and, at minimum, interval scaled person parameters. The Rasch model links the probability of a particular item response (e.g. correct) with the difficulty of the item, and the ability of the examinee. The model takes the form:

\[
P(\text{correct}) = \frac{e^{(\theta-b)}}{1+e^{(\theta-b)}} \tag{2}
\]

Where \( \theta \) is an individual’s ability and \( b \) is item difficulty with item discrimination assumed to be one. Discrimination is assumed to be one because the Rasch model assumes that an item discriminates perfectly between those with high ability and low ability.

\[
P(\text{correct}) = \frac{e^{(\theta-b)}}{1+e^{(\theta-b)}}
\]

The 2-parameter logistic model includes the individual’s ability, \( \theta \), the item discrimination, \( a \), and the item difficulty, \( b \):

\[
P(\text{correct}) = \frac{e^{1.7(a)(\theta-b)}}{1+e^{1.7(a)(\theta-b)}} \tag{4}
\]
This model includes a scaling constant, $D$, that is typically set to 1.7 in order to allow the model to be similar to the normal ogive and minimize the differences between the two functions (de Ayala, 2009).

The 3-parameter logistic model is expressed:

$$P \text{ (correct)} = c + (1 - c) \frac{e^{1.7(a)(\theta - b)}}{1 + e^{1.7(a)(\theta - b)}} \quad (5)$$

where $b$ is the item difficulty parameter, or the point at which an individual has an equal probability of getting an item correct or incorrect (Hambleton & Swaminathan, 1985). The item discrimination parameter, $a$, is the slope of the function at the point of inflection. The discrimination parameter tells us how well a specific item is able to distinguish between people of different ability levels (de Ayala, 2009). While ability level is typically used to define the discrimination parameter, this can be any latent trait of interest, including just possessing levels of a certain trait that is not necessarily higher or lower likelihood of answering an item ‘correct’ vs. ‘incorrect’ (such as political conservativeness). The $c$ parameter is a lower bound for the function, often referred to as the pseudo-guessing parameter. Guessing is affected by how many options the individual is given. For example, if there is a four-item question then guessing is set at .25. There are two general forms of guessing: blind and informed. Blind guessing means that the individual being testing has zero ideas about which item response might be a correct response. Informed guessing, on the other hand, occurs when an examinee has some information and are therefore making their guess based on partial knowledge and it is not a completely uninformed decision (Rogers, 1999). Martin, del Pino, and De Boek (2006) have stated that guessing parameters are a result of an interaction between guessing and person ability rather than just an
independent parameter. De Ayala has noted that the random guessing parameter is not reflected in the observed data (2009).

**Item characteristic curve (ICC)**

The ICC represents observed and expected values for a given item. Given an individual’s location, the probability of getting an expected value is plotted. Specific forms of relationships between the latent trait and an observed response are shown. Relating to monotonicity, if the assumption holds true then the ICC will be a linear graph. If the line is steep, there is better item discrimination whereas if the line is flatter there is less item discrimination.
Figure 1 demonstrates the Rasch model, which takes into account item difficulty and person ability level. It shows that item X1 is an easy item, as illustrated by the higher than average likelihood of very low ability persons getting the item correct. There is also evidence that the item is easier than others because in order to have more than a 0.5 probability of getting the item right the individual taking the assessment need only have slightly above an ability level of -2. It also demonstrates that the rest of the items all having discrimination that almost perfectly distinguishes between high and low ability levels.
Because the discrimination parameter is set to 1, this graph shows that the items are an appropriate difficulty as well.

Figure 2. ICC for 1PL

Figure 2 illustrates what happens when the item discrimination parameter goes from 1, as in the Rasch model, to being estimated from the data, as in the 1PL model. In this model, there is one discrimination parameter estimated for all items. We can see from these results that although the general patterns of the relationships between the latent trait and the probability of a correct response are similar to those in the Rasch model, the shapes of the ICCs have changed somewhat. This increase in item discrimination to a value greater than one means that for an item with this difficulty, the item discriminates between
high and low ability level individuals than was true for the Rasch, as evidenced by the steeper curve and some lower probability of lower levels correctly answering the item.

**Figure 3. ICC for 2PL**

Figure 3 shows what happens when item discrimination is estimated separately for each item. Most of the items retain a similar shape to the ICCs for the 1PL, but the ICC for item X0.2 becomes more linear than before, with an item discrimination value of .996. Item X1.1, however, becomes more vertical at about -1 for ability level (with a discrimination value of 2.023), showing that it discriminates very well for an item with an item difficulty level of -0.49.
Figure 4. ICC for 3PL

Figure 4 shows what the 3PL model, which includes guessing, looks like. Here, discrimination increases considerably for most items in comparison to the 2PL model. This is shown by the vertical points of the ICCs for three items, X1.1, X0, and X0.1. These are the points at which the item clearly discriminates, as for each small increase in ability, the probability of getting the item correct also increases. For example, between -1 and 1-ability levels, item X0 increases from about .2 to above .8 probability of getting the item correct. Item X1, however, is once again different than the other curves, demonstrating that with guessing, an item difficulty level of -.67, and an item discrimination value of 2.34, every person taking the assessment, regardless of ability level, has at least a 60% chance of
getting the item correct. With average ability, the individual is about 90% likely to get it
correct.

Assumptions of Item Response Theory

Item response theory has certain assumptions that must be met, including
unidimensionality, local independence, and monotonicity (deAyala, 2009). The first
assumption, unidimensionality, means that the latent trait that is of interest and is being
measured is one-dimensional. This means that only one trait is being measured. For
example, a test meant to measure math ability would not also measure reading ability. In
contrast to CTT, in which there is variability in the item parameters (difficulty and
discrimination) depending on who was in the sample, there is no variability and the sample
does not matter in creation of difficulty and discrimination because they should be the
same. Thus, the unidimensionality assumption makes IRT easier to use and more practical
than CTT, as the discrimination and difficulty parameters do not need to be estimated once
they are calculated for one sample. If the unidimensionality assumption is not met then
unidimensional IRT models cannot be used. However, multidimensional IRT models can be
used (Shealy & Stout, 1993).

The second assumption of IRT models is local independence. This assumption is
directly related to unidimensionality and states that item responses are not related to one
another, after conditioning on the latent trait. Thus, after controlling for ability level, $\theta$,
there are no correlations between test items. If a person answers one question a specific
way, that answer does not influence their response to another item on the assessment. If
there is local dependence, then how a person responds to one item will have an affect on
how a person responds to another item on the test. If this happens, then there is more than
one latent trait being measured, which is related to the first assumption of unidimensionality.

The third assumption is of IRT is monotonicity. This assumes that as an individual’s ability or knowledge increases so does the probability that the individual will give a correct response. For example, if a student took an item that was meant to be a measure of math ability then the student would be more likely to get the item right given a high level of \( \theta \) vs. low.

**Differential Item Functioning (DIF)**

Differential item functioning (DIF) is when two groups that took a test differ in ability to correctly answer a question on an assessment but have the same ability level. If two groups taking an assessment differ in how likely they are to have a particular response when the examinees have been matched on the latent trait of interest. For example, if math ability was the variable of interest and examinees had the same ability level, if one group was more likely than another to respond correctly to an item then DIF would be present. Differential item functioning (DIF) can be conceptualized in each of two ways. The first occurs when there are multiple dimensions. Standard IRT models have three assumptions, one of which is unidimensionality. This means that the model is measuring just one trait, rather than measuring more than one. If the assumption of unidimensionality is violated, there would be differences in probability of a correct response but that might be a result of a dimension that is not considered to be measured in the model. For example, if there is a math ability test but the questions are also worded such that they test reading ability, the assumption of unidimensionality is no longer met. This could mean that while two groups have the same math ability level, they have different overall reading levels that affect how
they are able to answer those questions that do measure math ability. Multi-dimensional models can be used, but more often unidimensional models are used for estimation and when that is the case, the DIF is caused by multidimensionality.

The second way that DIF can be conceptualized is if there are differences in item parameter values between the groups. If examinees differed on the IRT model parameters, that means that more than just one thing must be taken into account. For example, if one question is more difficult for a certain group of people more than another group, the difficulty parameter will vary depending on group. The item discrimination parameter must also be taken into account. If there is an item that better discriminates for gifted students, for example, better than typical students, the discrimination parameter must be taken into account.

Shealy and Stout (1993) propose that measuring more than one dimension with an item can be intentional or unintentional. If there are multiple dimensions being assessed by an item, the second dimension might cause one group to score differently than another as a result of nothing to do with the latent trait of interest. For example, if there is a math test that is written at a higher language level than necessary, a second latent trait (vocabulary) might also be being measured. If one of the groups is not as educated about the English language, they are at a disadvantage that has nothing to do with the latent trait of interest, math ability. As a result of the secondary dimension not being meant to be measured, a DIF analysis must be conducted to better understand exactly what the advantage that is being measured is.

The main concern if there is DIF is that the test or assessment will be unfair to certain group of people, even if they are matched with other individuals on ability level.
Test bias can be defined as being less valid for one group than another group and therefore is unfair in its attempt to assess examinee differences in a trait that was intended to be measured (Shealy & Stout, 1993). Validity is compromised because the test is supposed to measure the examinees ability on one specific latent trait. However, if there is bias, then the test is actually measuring more than just that and therefore is not a good measure of examinee’s ability on that one latent trait, but rather multiple latent traits or nuisance determinants. Item bias can be defined as a single item (versus the entire test) that is unfair in measurement of an intended trait. However, if an item is biased it only means that the item is answered differently by different groups regardless of ability level. Bias is inferred from DIF and typically is more conceptual in nature. DIF, on the other hand, means that there is a statistical difference, holding θ constant, in the likelihood of a group answering an item a specific way. It is important to note that just because there are group differences, does not necessarily mean that there is DIF. If the assessment is only measuring the dimension that it is intended to measure then it is valid and any group differences are due to knowledge of the construct being measured, not bias (Ackerman, 2005).

Because DIF is simply different item parameters for different groups of people, an IRF for each group on one plot can demonstrate DIF visually.
Figure 5. Uniform DIF, as ability level ($\theta$) increases, the probability ($p$) also increases.

Types of DIF

DIF can be either uniform or non-uniform. Uniform DIF can be explained as individuals in one group having a lower probability of correctly answering an item when matched at a given ability and holding that ability constant. In Figure 1, one group has a lower likelihood of answering the item correctly no matter what the $\theta$ level. For example, if a math assessment is administered to a group of males and a group of females, males might be less likely to answer the item correctly no matter than their ability level. If a male and female both had the ability level of 1, the female would be more likely to answer the question correct than the male, even though they have the same ability. In the figure above (Figure 1), that would mean that the males would be about 40% likely to answer correctly at an ability level of 1, but females would be about 80% likely on the same item with the same ability.
Non-uniform DIF can be defined as the discrimination parameter being different for each group. In Figure 2, non-uniform DIF can be seen where at certain \( \theta \) levels one group has a higher likelihood of answering an item correctly and at other \( \theta \) levels a different group has a higher likelihood of answering an item correctly. For example, if males and females were given the same math assessment as above the figure below (Figure 2) would demonstrate that males at an ability level of -1 would have a less than 10% chance of getting the item correct but females at the same ability level would have a 20% chance. However, as ability level increases to 1, males are more than 80% likely to answer the item correctly but females are now less likely than males, about 70%, to answer the item correctly. At the lower ability point, males are 10% LESS likely than females to get the item correct but with the higher ability point males are 10% MORE likely than females to get the item correct.

*Figure 6. Non-uniform DIF*
Methods for assessing DIF

There are currently a variety of statistical methods that have been suggested to detect DIF. Some of the methods are meant to compare pre-specified groups to one another using a focal and reference group and some, such as latent class approaches, allow DIF to be detected in groups that had exhibited it but were previously unknown. Some of the most popular methods include the Mantel-Haenszel (MH; Mantel & Haenszel, 1959), the logistic regression (LR) approach (Swaminathan & Rogers, 1990), the Thissen, Steinberg, and Wainer (1988) DIF detection strategy (IRTLR, the Item Response Theory Likelihood-Ratio Test (IRTLRT), SIBTEST (Shealy & Stout, 1993), & Rasch trees (Strobl, Kopf, & Zeileis, 2013).

Mantel-Haenszel

The Mantel-Haenszel statistic is a chi-square test of contingency that determines whether or not two variables are independent of one another and can be expressed:

\[ \psi_{MH} = \frac{\sum_i n_{00i}n_{11i}/n_i}{\sum_i n_{10i}n_{01i}/n_i} \]

\[ n_i = \sum_{x,y=0,1} n_{xyi} \] (6)

Where \( n_{xyi} \) represents the number of observations in a \((x,y)\) cell of the 2x2 table for a single, categorical variable. The MH statistic has multiple 2 x 2 chi-squares, with each representing individuals who received the same score on an assessment and shows the cross-classification of an item’s responses with group membership. This statistic allows for the responses to an item to be determined if they are independent of group membership after the observed scores are conditioned (de Ayala, 2009). Simply put, the MH statistic concludes the probability of the correct response for an item of a member of the focal
group versus the probability of a correct response for an item for a member of the reference group. An odds ratio across tables can be computed to determine the strength of the relationship.

**Logistic Regression**

Logistic regression (LR) makes predictions based on one or more quantitative or qualitative predictor variables about a binary variable. For example, the outcome variable would be correct response to an item and the predictor variable would be race. A logistic regression is conducted for one item using the reference group and then another is conducted for the same item with the focal group. If the two analyses differ in terms of intercepts then there is DIF present. However, if the two analyses have statistically same intercepts and regression coefficients, there is no evidence of DIF. In order to detect uniform DIF, the regression coefficients would be equal but the constants would not be equal. This would mean the probability curves are exactly the same, or parallel, but that they are not overlapping. Oppositely, if the coefficients were no equal then the probability curves would not be parallel (and would intercept one another) and there would be evidence of non-uniform DIF.

**Item Response Theory Likelihood Ratio**

Thissen, Steinberg, and Wainer’s DIF detection strategy compares two IRT models based on fit using the likelihood ratio test statistic (IRTLR; 1988). This test compares model fit takes the form:

\[ LR = -2 \ln L_c - (-2 \ln L_a) \] (7)

where \( L_c = \log \text{likelihood of a compact model} \) and \( L_a = \log \text{likelihood of the augmented model} \), when looking at a 2 parameter logistic model. The null hypothesis for the IRTLR
states that there are no group differences in item parameter estimates. The discrimination parameter may also be included in this test concurrently to see if this varies across groups as well with a 3PL model, with the addition of an allowance of variance between groups on the item difficulty parameter:

\[
LR_b = -2\ln L_c - (-2\ln L_{Ab}) \quad (8)
\]

where \( L_c \) is the log likelihood of the compact model and \( L_{Ab} \) is the log likelihood of the augmented model with the difficulty parameter being allowed to vary between groups. This strategy examines whether an item is constrained to one location across groups to the model fit if an item is allowed to move freely across groups. Thissen (2001) demonstrates the difference between the -2\( \ln L \) values for the models being presented as a chi-square statistic with 1 degree of freedom and therefore tests difficulty parameters and the presence of uniform DIF.

It is preferential that the location estimates for an item would not be significantly different between models, meaning that the item is performing identically in both the focal and reference group and that there is no evidence of DIF. If the models are significantly different and the item location varies between models then there is evidence of DIF, as the item performed differently for the focal versus reference group.

The calculations necessary to compute the test statistics in order perform IRTLR are very time-consuming. Studies have been done to show that the Type I error rates for the IRTLR test statistic was nominally higher (Cohen, Kim, & Wollack, 1996) and also that DIF might be susceptible to differences in Type I error based on percentage of DIF and underlying models (Atalay Kabasakal, Arsan, Gok, & Keleciolglu, 2014).
SIBTEST

SIBTEST is similar to the MH procedure in that bivariate frequencies are used to test frequencies of item responses and group memberships and is conditional on ability levels. SIBTEST is also a chi-square test and can be powerful in detecting unidirectional DIF, but can also detect non-unidirectional DIF. It does tend to perform better than MH and LR in terms of Type I error by controlling it and also reduces estimation bias because of a regression correction that is included in the procedure (Bolt, 2000; Gierl, Gotzmann, & Boughton, 2004). SIBTEST has been shown to be similar to MH in both Power and Type I error rate and fairly effective (Narayanan & Swaminathan, 1994; Navas – Ara & Gomez – Benito, 2002; Swaminathan & Rogers, 1993).

While the IRTLR approach to detecting DIF is IRT based, the LR and MH are not. There is a potential advantage to using the LR approach compared to the TSW-ΔG² test, however, in the sense that the LR approach does not hold any assumptions of IRT be met in order to use the test properly. Another possible positive to using the LR DIF detection test is that it can be used to determine if there is uniform DIF and well as non-uniform DIF, while the MH technique might or might not detect and identify non-uniform DIF, depending on cancellation (Rogers & Swaminathan, 1993; Narayanan & Swaminathan, 1996; de Ayala, 2009). The MH also does not allow for non-discrete person location predictors because the observed scores are used for conditioning while the LR analysis uses other variables rather than the observed scores and allows for covariates. Overall, literature is mixed about which DIF detection strategy is best and is dependent on conditions. Some of the tests currently used have a Type I error rate affected by sample size, models tested, DIF percentage, test
length, ability differences, or interactions of these different variables (Atalay Kabasakal, Arsan, Gok, Kelecioglu, 2014).

SIBTEST seems to exhibit a significantly higher Type I error rate than both the Mantel-Haentzel and the IRT-LR. Specifically, SIBTEST seems to be affected by ability distribution, such that focal and reference groups with similar standard deviations showed a higher Type I error rate. SIBTEST also did better with longer test than shorter, with TYPE I error rate being lower for tests with more questions. SIBTEST also had trouble with the model type, performing better on Type I error with 2PL models rather than three. IRT-LR performed opposite, with Type I error being inflated in a 2PL model rather than a 3PL.

Mantel-Haentzel seems to be affected in Type I error rate by sample size, test length, ratio of DIF, and model type. With equal size focal and references groups, higher Type I error seems to be exhibited. Type I error rate also seems to be higher when MH is used for assessments with more questions and when used in simulation studies with smaller amounts of DIF, about 5% (Atalay Kabasakal et al., 2014).

*Rasch Tree*

Recently, a new semi-parametric approach of testing for DIF in Rasch Models, based on recursive partitioning, has been proposed by Strobl, Kopf, & Zeileis (2013). Semi-parametric approaches simply have components of both paramateric and non-parametric approaches. In this case, the non-parametric component is that the recursive partitioning is not assumed to be linear. Using the Rasch Tree Method, defined covariates, or combinations of covariates, would be used to detect differences in model parameters. Using modern statistical techniques in order to be more robust than classification and regression trees (CART), the current approach uses parametric model parameters in place of one
response variable. Item responses are recorded taking into account group differences related to covariates. The partitioned model is then presented in the form of a figure, or tree, with each terminal node displaying parameter estimates (see figure 3.). If there is more than one node, then the null hypothesis being tested, that there are no group differences, must be rejected. One advantage to Rasch Trees is that the results are presented visually in a way that makes it easy to see which groups are affected by DIF in regards to specific items (Strobl, Kopf, & Zeileis). This is more exploratory than some other methods, which is a positive feature of this model in the sense that variables need not be specified to the algorithm and that makes it more flexible than parametric methods in detecting groups affected by DIF. In other words, where other models require the researcher know what variables exhibit DIF before fitting the model to the data, the current method does not require knowledge of what variables will exhibit DIF before running the model. RaschTree will automatically cut the data based on stability levels and finally stop splitting when the data have reached the most stable conditions, which means that there is no longer DIF exhibited.

The method is grounded in model based recursive partitioning related to classification and regression trees (CART), employing tests for structural change in a semi-parametric approach. Covariates of groups of subjects are analyzed to detect differences in parameters of statistical models. The algorithm used works as follows:

1. Estimate Rasch model parameters jointly for all persons in the full sample.
2. Assess stability of item parameters with respect to the DIF variable(s) of interest (e.g. gender, ethnicity, age).
3. If parameters are unstable, split sample for the DIF variable with the greatest instability using the cutpoint of the model parameter that leads to the greatest stability possible.

4. Repeat steps 1-3 recursively within the subsamples until all model parameters are stable.

Figure 7. Example of Rasch Tree

The Rasch tree method offers an advantage over other methods by combining automatic detection of groups exhibiting DIF while at the same time having covariate
values that are interpretable. A cutpoint does not have to be specified beforehand for focal and reference groups. The strongest parameter difference, instead, has an affiliated cutpoint that is automatically detected. This means that by defining groups less than optimal does not have such a negative effect because DIF cannot go unnoticed for a numeric covariate. Another advantage that the Rasch tree has over other methods is that because of its graphical nature, both groups of items and groups of subjects affected by additional dimensions can be easily identified (Strobl, Kopf, Zeileis, 2010).

While RaschTree does have the ability to explore DIF that other methods do not, one major short-coming of RaschTree is that it is unable to tell you where exactly the DIF is located. For example, which RaschTree may be able to detect DIF in variables such as gender, it cannot tell you exactly which items. In order to look at that, another method would have to be used. However, this could be a method that is able to exploratorily find DIF that was not even considered before, then those variables can be put through another test that can pinpoint the items to better fix any DIF issues.

Another issue with RaschTree is that while breaking down the groups that are exhibiting DIF, eventually there are terminal groups that have very small samples. Small samples can often have Type II error skewing results, which decreases power.

Figure 3 demonstrates a Rasch tree that depicts gender exhibiting DIF. In the figure, you can see that from the gender variable there are two nodes, one representing male and one for female. From each of those nodes, there are two nodes (2 and 5) that are further broken down to more specific DIF groups until a terminal node (3, 4, 6, and 7) is reached. These terminal nodes represent estimates of the item difficulty for each item for the combination of nodes that lead to it.
Comparison of methods

The Mantel-Haenszel (MH) statistic (Mantel & Haenszel, 1959) has been shown to have inflated Type I error rate if DIF is non-existent (Magis & DeBoeck, 2012; Wang & Su, 2004). It also cannot detect non-uniform DIF. However, it is used more often than some other DIF detection strategies as a result of having fewer disadvantages compared to advantages (Millsap & Everson, 1993; Thissen, 2001). MH performs well with smaller sample sizes of about 200 to 250 people per group (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993) that do not meet requirements for other DIF detection strategies, namely IRT-LRT and LR (Narayanan & Swaminathan, 1994) and has been shown to have consistent power even with unequal group ability distributions (Guler & Penfield, 2009). It also provides \( \Delta \), an effect size, along with the statistical test. These advantages, along with the inexpensive implementation make it an often used DIF test (French & Finch, 2013).

Logistic Regression (LR) DIF detection (Swaminathan & Rogers, 1990) can be used to detect both uniform and non-uniform DIF in dichotomous and polytomous data with comparable power to MH when detecting unidirectional DIF and is actually more powerful than MH when detecting non-unidirectional DIF (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993). However, inflated Type I error rates have been shown under the condition of unequal group ability distribution (Guler & Penfield, 2009), and it is more computationally rigorous because of the necessary iterative parameter estimation (Guler & Penfield, 2009; Narayanan & Swaminathan, 1996; Swaminathan & Rogers, 1990).
The Thissen, Steinberg, and Wainer (1988) DIF detection strategy, the Item Response Theory Likelihood-Ratio Test (IRTLRT), can be used for dichotomous data and polytomous data and holds some advantages over other DIF tests. It is a powerful test and can evaluate items in more than two groups (Sireci, Patsula, & Hambleton, 2005), shows adequate Type I error control for both dichotomously and polytomously-scored items using the 2PL (Cohen et al., 1996) and the graded response model (Kim & Cohen, 1998), respectively. While IRTLRT does have several advantages, one major disadvantage is that at this time it is necessary to fit multiple IRT models to data. Also, assessments with a large number of items with no rejection of a DIF model make it difficult to figure out which items have DIF (Sireci, Patsula, & Hambleton). IRTLRT has also been shown to provide adequate Type I error control across sample sizes and distribution conditions in computerized adaptive testing (Lei, Chen, & Yu, 2006).

SIBTEST (Shealy & Stout, 1993) is able to detect uniform and non-uniform DIF in dichotomous data and is at an advantage of other detection strategies that can detect both because it can account for ability level differences, is still effective with small sample sizes, has an effect size estimate, and can be used to assess differential bundle functioning (Lee, Cohen, & Toro, 2009; Lei & Li, 2013; Pei & Li, 2010; Roussos & Stouth, 1996; Walker, Zhang, & Banks, 2012) SIBTEST has also performed well in comparative studies with simulations of Type I error compared to other DIF detection strategies, namely MH and LR (Bolt, 2000; Li & Stout, 1996; Roussos & Stout, 1996) and short tests and large samples (DeMars, 2009). Even with overall good performance, a substantially inflated Type I error rate has been shown in a small number of cases. However, even in these minority cases, other DIF
detection strategies show higher Type I error rate (Roussos & Stout, Shih, Liu, & Wang, 2014).

Previously mentioned DIF detection strategies estimate parameters between groups that are pre-defined, for example, men and women (Strobl, Kopf, & Zeileis, 2010). In this example, men would be the reference group and women would be the focal group. In these scenarios, DIF detection tells us which group has lower scores and the test can be reevaluated to figure out how to solve that problem. However, the main disadvantage of these detection strategies that the Raschtree Method seeks to remedy is that the groups must be predefined. The researchers must be able to know which group to assess for DIF and does not allow for non-predefined group differences to be detected or explained.

Latent class analyses do allow for DIF detection of groups that are non-predefined but are not directly interpretable. Because Raschtree allows for detection of both non-predefined groups as well as latent groups and is still interpretable, it is more useful than previous methods. Previous simulation studies done over multiple conditions and using other DIF detection strategies has given evidence that a more efficient test to detect DIF while minimizing Type I error and increasing power needs to be put into use. While other tests do exist that can detect DIF in data, the conditions under which they all work individually are limited, as previously discussed. Rasch tree is meant as a solution to fill that gap and be able to detect DIF and estimate item difficulty parameters in more than just dichotomous data, namely trichotomously and polytomously scored data, while still controlling Type I error and power.

RaschTree was originally presented as a new method to detect DIF by combining latent class analysis and advantages of previous approaches for given groups. Groups that
do have DIF are interpretable but also are not pre-specified, which is something that other methods were either unable to do or unable to do in a way that was easily interpretable (Strobl, Kopf, & Zeileis, 2013). Because the cutpoint associated with the largest instability in parameters is automatically detected, a covariate with exhibiting DIF could not go unidentified due the large instability between models. This being said, if there are multiple variables that exhibit DIF included in the Tree, theoretically RaschTree should not be able to miss any of the instability splits and therefore detect DIF in multiple covariates, as well as interactions, at once.

One thing that RaschTree cannot do, which is a shortcoming of any covariate-based approach, is detect covariates that were not identified as necessary to be included in the model. If there is a covariate of interest that may exhibit DIF, of course the RaschTree approach cannot detect it, as it is not there. While DIF may be detected in a related variable, that other variable would then have to be actually put in the model to find the instability points to look at if DIF is present. It is also noted by Strobl et al. (2013) that although DIF can be detected in variables, those variables cannot be said for certain to be causal. For example, if there is a split between males and females, that may be a larger issue behind the resources and attention that are given to each gender in different settings.

The current study seeks to address the following research questions:

(1) What results does RaschTree produce when used to detect DIF for multiple variables at once?

(2) Is Rasch Tree able to detect interactions between variables in DIF?

(3) Are there any benefits that RaschTree provides in detecting DIF that other detection methods do not?
CHAPTER 3

METHOD

In order to address the research questions of interest, namely investigation of RaschTree in detecting DIF in multiple variables at once as well as interactions between variables, the Progress in International Reading Literacy Study (PIRLS) data (2011) was analyzed. The PIRLS is an assessment that has been administered every 5 years (beginning in 2001) and measures and compares student learning in reading. Fourth-grade students are asked to complete a reading assessment and questionnaire that addresses the student’s attitudes and habits towards reading. The purpose of the PIRLS is to track and help understand trends in reading knowledge of 4th graders as well as help document instruction practices used by instructors. The PIRLS provides researcher with the ability to childhood literacy world-wide, which in turn provides policy makers with information necessary to make important decisions regarding reading education. It also yields insights into how much 4th grade students attitudes and enjoyment in reading vary, and how they might influence reading performance. Therefore, if there is DIF in the PIRLS items, incorrect decisions regarding about education systems and/or individuals may be made.

Questionnaires were administered to students, teachers, and principals. The questionnaires administered to students, which are included in results of this study, included both a reading assessment based on a short story as well as a scale that measured attitudes towards reading. The first year of administration, in 2001, 36 education systems were given a reading assessment and questionnaire; five years later a second administration was given (2006) to 45 education systems; five years later (2011), the PIRLS was administered to 53 education systems. ‘Education system’ was used as a
descriptor in order to include countries as well as subnational entities such as provinces and administrative regions.

There were multiple variables included in the current study: gender, a scale that measured attitudes toward reading (including enjoyment, motivation, confidence, and engagement), and correct or incorrect answers on a reading achievement scale, which was a 13-question scale based off a short story.

**Gender**

Gender was coded as male or female, with some students failing to report their gender. Gender was used as the dichotomous variable in this study at which point RaschTree was used to find possible DIF between the two variables.

**Attitude toward reading scale**

The attitude towards reading scale was administered to students, with the students themselves rating how much they enjoyed reading, how motivated they were to read, how much confidence they had about their reading skills, and how engaged they were when reading. This scale was a three-point ordinal scale ranging from 1 to 3 with 1 being positive (such as ‘like reading’, ‘motivated’, ‘confident’, and ‘engaged’) and 3 being negative (such as ‘do not like reading’, ‘not motivated’, ‘not confident’, and ‘not engaged’).

**PIRLS questions**

Responses to PIRLS items were included as the outcome variables that may reflect DIF. The PIRLS subtest included in this study was an 811-word-count passage with a Flesch-Kincaid grade level of 2.8. The Flesch-Kincaid Grade Level Formula uses average syllables per word and average sentence length to produce a number that represents the US grade in
which students can read the text.

After reading the text, students were asked to answer 13 questions to test their comprehension and literacy of the passage. These questions took the form of both multiple choice and short-answer. While some items were scores on an ordinal scale (e.g. no comprehension, partial comprehension, total comprehension), the coding of the dataset did have a correct and incorrect code as well that was done by scorers. This is what was used in this study.

After cleaning the dataset up for missing data in the PIRLS items and gender, RaschTree analysis was used to detect DIF in the PIRLS items for both gender and attitudes toward reading as well as the interaction of DIF present in both using the package psychotree (Zeileis, Strobl, Wickelmaier, & Kopf, 2010) in R (R Development Core Team, 2010).

**Settings**

For this specific RaschTree application, default settings of the Raschtree were implemented. This being that the nodes of the tree, ‘coef’ extracts all item parameters except the first, which is always restricted to zero. ‘Itempar’ extracts all item parameters and by default restricts their sum to zero, and the ‘plot’ method by default employs the panel-generating function ‘node_profileplot’ and the panel-generating ‘node_regionplot’ as an alternative.

A primary purpose of this study was to explore how RaschTree reacted to attempting to measure DIF in multiple variables without a set cutpoint as well as to examine if RaschTree is even able to detect DIF when interactions are included in the exploratory model.
CHAPTER 4

RESULTS

The data collected was designed to assess and compare reading achievement among school-aged children internationally. PIRLS data was collected in the United States by the National Center for Education Statistics (NCES) and was sponsored by the International Association of the Evaluation of Educational Achievement (IEA). The PIRLS included 10,013 students, with 51.6% (n=5169) being male and 47.5% (n=4761) being female. Eighty-three students failed to report their gender. Table 1 shows correct and incorrect responses to items included on the scale used to assess reading comprehension.
Table 1

Percentage Correct for PIRLS Reading Comprehension Items

<table>
<thead>
<tr>
<th>PIRLS Item</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>8288</td>
<td>1725</td>
<td>82.77%</td>
</tr>
<tr>
<td>Male</td>
<td>4099</td>
<td>1070</td>
<td>79.30%</td>
</tr>
<tr>
<td>Female</td>
<td>4127</td>
<td>634</td>
<td>86.68%</td>
</tr>
<tr>
<td>Item 2</td>
<td>8237</td>
<td>1776</td>
<td>82.26%</td>
</tr>
<tr>
<td>Male</td>
<td>4237</td>
<td>932</td>
<td>81.97%</td>
</tr>
<tr>
<td>Female</td>
<td>4036</td>
<td>725</td>
<td>84.77%</td>
</tr>
<tr>
<td>Item 3</td>
<td>7765</td>
<td>2248</td>
<td>77.55%</td>
</tr>
<tr>
<td>Male</td>
<td>3905</td>
<td>1264</td>
<td>75.55%</td>
</tr>
<tr>
<td>Female</td>
<td>3910</td>
<td>851</td>
<td>82.13%</td>
</tr>
<tr>
<td>Item 4</td>
<td>8902</td>
<td>1111</td>
<td>88.90%</td>
</tr>
<tr>
<td>Male</td>
<td>4532</td>
<td>637</td>
<td>87.68%</td>
</tr>
<tr>
<td>Female</td>
<td>4369</td>
<td>392</td>
<td>91.77%</td>
</tr>
<tr>
<td>Item 5</td>
<td>7957</td>
<td>2056</td>
<td>79.50%</td>
</tr>
<tr>
<td>Male</td>
<td>4026</td>
<td>1143</td>
<td>77.89%</td>
</tr>
<tr>
<td>Female</td>
<td>4001</td>
<td>760</td>
<td>84.04%</td>
</tr>
<tr>
<td>Item 6</td>
<td>6464</td>
<td>3549</td>
<td>64.56%</td>
</tr>
<tr>
<td>Male</td>
<td>3465</td>
<td>1702</td>
<td>67.03%</td>
</tr>
<tr>
<td>Female</td>
<td>3170</td>
<td>1590</td>
<td>66.58%</td>
</tr>
<tr>
<td>Item 7</td>
<td>3469</td>
<td>6544</td>
<td>34.64%</td>
</tr>
<tr>
<td>Male</td>
<td>1797</td>
<td>3370</td>
<td>34.76%</td>
</tr>
<tr>
<td>Female</td>
<td>2005</td>
<td>2755</td>
<td>42.11%</td>
</tr>
<tr>
<td>Item 8</td>
<td>6909</td>
<td>3104</td>
<td>69.00%</td>
</tr>
<tr>
<td>Male</td>
<td>3433</td>
<td>1735</td>
<td>66.42%</td>
</tr>
<tr>
<td>Female</td>
<td>3595</td>
<td>1166</td>
<td>75.51%</td>
</tr>
<tr>
<td>Item 9</td>
<td>7256</td>
<td>2757</td>
<td>72.47%</td>
</tr>
<tr>
<td>Male</td>
<td>3654</td>
<td>1503</td>
<td>70.70%</td>
</tr>
<tr>
<td>Female</td>
<td>3686</td>
<td>1070</td>
<td>77.42%</td>
</tr>
<tr>
<td>Item 10</td>
<td>8718</td>
<td>1295</td>
<td>87.07%</td>
</tr>
<tr>
<td>Male</td>
<td>4460</td>
<td>709</td>
<td>86.28%</td>
</tr>
<tr>
<td>Female</td>
<td>4279</td>
<td>482</td>
<td>89.88%</td>
</tr>
<tr>
<td>Item 11</td>
<td>5939</td>
<td>4074</td>
<td>59.31%</td>
</tr>
<tr>
<td>Male</td>
<td>3210</td>
<td>1959</td>
<td>62.10%</td>
</tr>
<tr>
<td>Female</td>
<td>2922</td>
<td>1839</td>
<td>61.37%</td>
</tr>
<tr>
<td>Item 12</td>
<td>2247</td>
<td>7766</td>
<td>22.44%</td>
</tr>
<tr>
<td>Male</td>
<td>1159</td>
<td>4010</td>
<td>22.42%</td>
</tr>
<tr>
<td>Female</td>
<td>1514</td>
<td>3247</td>
<td>31.80%</td>
</tr>
<tr>
<td>Item 13</td>
<td>7072</td>
<td>2941</td>
<td>70.63%</td>
</tr>
<tr>
<td>Male</td>
<td>3614</td>
<td>1555</td>
<td>69.92%</td>
</tr>
<tr>
<td>Female</td>
<td>3587</td>
<td>1173</td>
<td>75.34%</td>
</tr>
</tbody>
</table>
Responses for reading enjoyment, motivation to read, confidence in reading, and engagement in reading were also recorded, with responses ranging from 1 to 3, with 1 being positive (such as 'like reading', 'motivated', 'confident', and 'engaged') and 3 being negative (such as 'do not like reading', 'not motivated', 'not confident', and 'not engaged'). Means for these ordinal variables are shown in Table 2.

Table 2

<table>
<thead>
<tr>
<th>Attitude Scale</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student likes reading</td>
<td>1.89</td>
<td>0.66</td>
</tr>
<tr>
<td>Male</td>
<td>2.03</td>
<td>0.65</td>
</tr>
<tr>
<td>Female</td>
<td>1.74</td>
<td>0.62</td>
</tr>
<tr>
<td>Student motivated to read</td>
<td>1.30</td>
<td>0.53</td>
</tr>
<tr>
<td>Male</td>
<td>1.34</td>
<td>0.57</td>
</tr>
<tr>
<td>Female</td>
<td>1.25</td>
<td>0.48</td>
</tr>
<tr>
<td>Student confident in reading</td>
<td>1.72</td>
<td>0.64</td>
</tr>
<tr>
<td>Male</td>
<td>1.78</td>
<td>0.65</td>
</tr>
<tr>
<td>Female</td>
<td>1.67</td>
<td>0.63</td>
</tr>
<tr>
<td>Student engaged in reading</td>
<td>1.67</td>
<td>0.60</td>
</tr>
<tr>
<td>Male</td>
<td>1.75</td>
<td>0.62</td>
</tr>
<tr>
<td>Female</td>
<td>1.58</td>
<td>0.57</td>
</tr>
</tbody>
</table>

RaschTree for Gender

Figure 1 shows the Rasch Tree for gender, with males and females answering significantly different \( (p < .001) \) to items on the PIRLS assessment, meaning there is differential item functioning (DIF) detected.
Figure 8. Rasch Tree for Gender
As shown by the Rasch Tree in Figure 1, the split between the gender variable illustrates significant differential item functioning between the genders. Looking at Figure 1, it is clear that there is a larger difference in difficulty points on the plot in items 6 and 11. Table 3 more comprehensively shows what items showed DIF and what exactly the item difficulty parameters were. For item six, ‘granny screamed’, females were more likely to correctly respond to the item after reading the passage than males, after controlling for the latent trait being measured, in this case reading comprehension. For item 11, ‘purpose of the last line of the story’, females are again more likely to respond to the item correctly, when controlling for the latent trait being measured.

Table 3

<table>
<thead>
<tr>
<th>Item Difficulty Parameters</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>-1.03</td>
<td>-0.72</td>
</tr>
<tr>
<td>Item 2</td>
<td>-0.57</td>
<td>-0.78</td>
</tr>
<tr>
<td>Item 3</td>
<td>-0.47</td>
<td>-0.35</td>
</tr>
<tr>
<td>Item 4</td>
<td>-1.48</td>
<td>-1.42</td>
</tr>
<tr>
<td>Item 5</td>
<td>-0.83</td>
<td>-0.67</td>
</tr>
<tr>
<td>Item 6</td>
<td>0.63</td>
<td>0.20</td>
</tr>
<tr>
<td>Item 7</td>
<td>1.77</td>
<td>1.80</td>
</tr>
<tr>
<td>Item 8</td>
<td>0.12</td>
<td>0.21</td>
</tr>
<tr>
<td>Item 9</td>
<td>-.07</td>
<td>-.08</td>
</tr>
<tr>
<td>Item 10</td>
<td>-1.29</td>
<td>-1.27</td>
</tr>
<tr>
<td>Item 11</td>
<td>0.81</td>
<td>0.47</td>
</tr>
<tr>
<td>Item 12</td>
<td>2.34</td>
<td>2.60</td>
</tr>
<tr>
<td>Item 13</td>
<td>0.07</td>
<td>0.01</td>
</tr>
</tbody>
</table>

*RaschTree for Reading Engagement*

Figure 2 illustrates the Rasch Tree for the variables of enjoyment, motivation, confidence, and engagement in reading. The first split in the tree was for confidence, with students who were not confident splitting from those who were somewhat or totally
confidant. The second split in the tree was for reading enjoyment, with those who like reading splitting from those who somewhat like reading and who do not like reading. As illustrated, the Rasch tree has splits for the variables of confidence in reading and enjoyment in reading, indicating DIF in these particular variables but not for the variables motivation to read or engagement in reading. Figure 2 also demonstrates that not only is not being confident a determining factor of how hard a question may be to a student, but also an interaction between confidence in reading (if they are somewhat or totally confident) and enjoyment that determines which items are more difficult of easy to solve.
Figure 9. Rasch Tree illustrating DIF for continuous variables
Table 4

**Item Difficulty Parameters of Ordinal Variables**

<table>
<thead>
<tr>
<th>PIRLS Item</th>
<th>Confident, Somewhat Confident, Like Reading</th>
<th>Confident, Somewhat Like Reading, Do not Like Reading</th>
<th>Not Confident</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>-1.04</td>
<td>-0.83</td>
<td>-0.68</td>
</tr>
<tr>
<td>Item 2</td>
<td>-0.54</td>
<td>-0.69</td>
<td>-0.91</td>
</tr>
<tr>
<td>Item 3</td>
<td>-0.31</td>
<td>-0.39</td>
<td>-0.69</td>
</tr>
<tr>
<td>Item 4</td>
<td>-1.44</td>
<td>-1.47</td>
<td>-1.36</td>
</tr>
<tr>
<td>Item 5</td>
<td>-0.83</td>
<td>-0.71</td>
<td>-0.73</td>
</tr>
<tr>
<td>Item 6</td>
<td>0.50</td>
<td>0.40</td>
<td>0.25</td>
</tr>
<tr>
<td>Item 7</td>
<td>1.76</td>
<td>1.82</td>
<td>1.62</td>
</tr>
<tr>
<td>Item 8</td>
<td>0.07</td>
<td>0.15</td>
<td>0.44</td>
</tr>
<tr>
<td>Item 9</td>
<td>-0.15</td>
<td>-0.10</td>
<td>0.17</td>
</tr>
<tr>
<td>Item 10</td>
<td>-1.27</td>
<td>-1.27</td>
<td>-1.30</td>
</tr>
<tr>
<td>Item 11</td>
<td>0.85</td>
<td>0.60</td>
<td>0.33</td>
</tr>
<tr>
<td>Item 12</td>
<td>2.35</td>
<td>2.51</td>
<td>2.56</td>
</tr>
<tr>
<td>Item 13</td>
<td>0.06</td>
<td>-0.02</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table 4 shows the difficulty parameters of each node. From Table 4, it can be seen that items 1 (need to look at key), 2 (‘granny’s farmhouse’), 3 (‘reason granny should move’), item 8 (‘granny winked and grinned’), and item 11 (‘purpose of last line of the story’) all showed very large differences between how groups answered, with the largest difference being between confidence levels, with no confidence being the main split. After the split between not confident and somewhat and very confident readers, enjoyment of reading was also a factor.

As discussed previously, item difficulty parameter can be explained as a $b$ value that corresponds to the ability level at which a question has a 50% chance of being answered correctly. As can be seen above, items 7 and 12, in particular, were
more difficult for all groups, with ability having to be over 1 and 2 respectively to have an equal chance to answer those questions correctly. Interestingly, students who were ‘not confident’ with no other factors involved (the main split in the Rasch Tree), sometimes found questions easier than their counterparts who were ‘confident’ or ‘somewhat confident’. For example, items 2, 3, 6, 7, and 11 had the lowest difficulty level for ‘not confident’ students. Overall, the group who rated themselves as at least somewhat confident and liked reading had the lowest item difficulty values. That group had 6 questions that they had the lowest item difficulty level for (questions 1, 5, 8, 9, 10, and 12). Group 2, those who were at least somewhat confident and either somewhat or did not like reading had the lowest difficulty level on only three questions (items 4, 10, and 13). Values between the groups with the lowest difficulty level and the highest difficulty level ranged from a difference of .52 to only .03.

**Rasch Tree for All Variables**

Figure 3 shows the Rasch Tree for all variables, including gender, confidence in reading, motivation to read, engagement in reading, and enjoyment in reading. This model differs from previously presented models in that it contains multiple continuous covariates, thereby allowing for more complex patterns of DIF to be detected. Indeed, as illustrated in the results, DIF was detected for multiple items, with interactions among multiple variables being associated with this finding. For boys, confidence was a significant split ($p < .001$), with boys who are not confident ($n = 395$) responding to questions differently than boys who are either somewhat or
totally confident in their reading ability (n = 2844). For girls, both confidence ($p < .05$) and engagement ($p < .05$) played a role in likelihood of correctly answering.
Figure 10. Rasch Tree illustrating DIF for all variables
Table 5

*Combined Variable Item Difficulties*

<table>
<thead>
<tr>
<th>PIRLS Item</th>
<th>Female, Confident, Somewhat Confident, Engaged</th>
<th>Female, Confident, Somewhat Confident, Not Engaged</th>
<th>Female, Not Confident</th>
<th>Male, Confident, Somewhat Confident</th>
<th>Male, Not Confident</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>-1.02</td>
<td>-1.06</td>
<td>-0.96</td>
<td>-0.76</td>
<td>-0.53</td>
</tr>
<tr>
<td>Item 2</td>
<td>-0.43</td>
<td>-0.63</td>
<td>-0.90</td>
<td>-0.75</td>
<td>-0.92</td>
</tr>
<tr>
<td>Item 3</td>
<td>-0.45</td>
<td>-0.40</td>
<td>-0.84</td>
<td>-0.31</td>
<td>-0.61</td>
</tr>
<tr>
<td>Item 4</td>
<td>-1.53</td>
<td>-1.51</td>
<td>-1.23</td>
<td>-1.42</td>
<td>-1.43</td>
</tr>
<tr>
<td>Item 5</td>
<td>-0.97</td>
<td>-0.74</td>
<td>-0.61</td>
<td>-0.65</td>
<td>-0.81</td>
</tr>
<tr>
<td>Item 6</td>
<td>0.70</td>
<td>0.60</td>
<td>0.43</td>
<td>0.21</td>
<td>0.13</td>
</tr>
<tr>
<td>Item 7</td>
<td>1.88</td>
<td>1.69</td>
<td>1.71</td>
<td>1.83</td>
<td>1.56</td>
</tr>
<tr>
<td>Item 8</td>
<td>0.06</td>
<td>0.11</td>
<td>0.43</td>
<td>0.18</td>
<td>0.44</td>
</tr>
<tr>
<td>Item 9</td>
<td>-0.06</td>
<td>0.11</td>
<td>0.06</td>
<td>-0.13</td>
<td>0.24</td>
</tr>
<tr>
<td>Item 10</td>
<td>-1.39</td>
<td>-1.22</td>
<td>-1.20</td>
<td>-1.25</td>
<td>-1.36</td>
</tr>
<tr>
<td>Item 11</td>
<td>0.89</td>
<td>0.80</td>
<td>0.47</td>
<td>0.50</td>
<td>0.24</td>
</tr>
<tr>
<td>Item 12</td>
<td>2.22</td>
<td>2.48</td>
<td>2.37</td>
<td>2.59</td>
<td>2.74</td>
</tr>
<tr>
<td>Item 13</td>
<td>0.10</td>
<td>0.00</td>
<td>0.26</td>
<td>-0.04</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Table 5 shows the item difficulty values for all of the variables, including gender, confidence level, engagement, enjoyment, and motivation. As shown in Figure 3, there was significant DIF detected, with splits first forming at gender, followed by confidence, than engagement. After the split between male and female, confidence split between ‘not confident’ and other levels of confidence for both genders. For
girls, there was then another split between being engaged, and being somewhat or not engaged. Significant differences in terminal nodes can be identified in Table 5 along with item difficulty parameters. As illustrated, specifically interesting is item 1 (identity of narrator), which illustrates an almost .5 difference in item difficulty between different nodes (with the largest gap being between engaged, confident females and non-confident males). Item 3 (‘reason granny should move’) demonstrated a .39 difference in item difficulty between engaged, confident females and non-confident females. Finally, item 13 (‘lesson learned from story’), demonstrated both positive and negative item difficulty values, demonstrating a .39 difference between males who were somewhat or not totally confident and males who were not confident.

Comparing the three models, it can be seen that splits between the variables were found at all levels. Rasch Tree was able to detect DIF in all three models, with the more complex still having splits among the groups. In looking closely at the terminal nodes and item difficulties, it can be seen that although the significance levels of the more complex models are not as small as the significance levels for the simpler models that involve only one variable, they still fit well within the typical level used in this type of testing and assessment (p < .05).
CHAPTER 5
DISCUSSION

A variety of negative outcomes can result from the presence of DIF in test questions. These consequences can include decisions made by educators regarding academic placement, provision of services, and student access to special programs. For example, the DIF detected between gender could mean that males are put into remedial classrooms when, in fact, it is not their ability level that is lacking but that males are not being fairly tested with reading assessments. As an example regarding this study in terms of DIF with multiple ordinal variables, the group that is at least somewhat confident and likes reading is more likely to answer the assessment questions correctly, so therefore may be given opportunities to be enrolled in gifted programs and therefore set up to succeed later in school more so than the groups that are not confident, but that is due to the assessment measuring confident and enjoyment rather than actual comprehension. As far as interactions that were found to exhibit DIF, certain groups would again, be benefitting from the way the assessment is written and may just possess a combination of traits rather than reading skills. For example, a not confident female may score lower than a confident male, again, based on the traits of gender and confidence rather than actual reading ability. Any of the traits that are being measured by an assessment put not only school systems and decisions regarding those score in jeopardy, but also a much larger impact on the individual who will deal with both the short-term and long-term affects of unfair assessments.
Traditional methods for DIF detection focus on one variable at a time, and typically work best with discrete categorical variables. Despite the fact that such methods have been shown to be effective in detecting DIF in such relatively simple situations (MH; Mantel & Haenszel, 1959, the logistic regression (LR) approach (Swaminathan & Rogers, 1990), the Thissen, Steinberg, and Wainer (1988) DIF detection strategy (IRTLR, the Item Response Theory Likelihood-Ratio Test (IRTLRT), SIBTEST (Shealy & Stout, 1993), a method that could incorporate multiple variables simultaneously would be useful in identifying the multifaceted nature of test performance. RaschTree appears to offer just such an option for bridging the gap between detecting DIF for multiple variables simultaneously and the limitations of previously proposed methods, specifically to be able to detect DIF in trichotomous and polytomous data well.

The goal of this study was to demonstrate the utility of RaschTree for detecting DIF, and to investigate how it can be employed with multiple variables simultaneously in DIF detection. Influenced by a model based on semi-parametric approach based on recursive partitioning (Strobl, Kopf, & Zeileis, 2010), the current study sought to answer questions of demonstrating RaschTree’s utility for different variable types, including multiple variables, interaction between variables, a visual demonstration of what was found, and if it would work to find cutpoints that were not pre-specified, but rather an exploratory approach.

Study Findings

Currently, Classical Test Theory and Item Response Theory are often used in educational assessment and measurement. IRT focuses on individual item
performance versus CTT, which focuses on a total score that includes only the true score plus error. There are a variety of latent class approaches used to detect DIF, but there is no definitive evidence that one method works best. Rather, the most effective technique for a given situation is dependent on conditions such as number of items, sample size, and number of groups. These factors can inflate the Type I error rate of such methods and thereby potentially limit their utility. In order to remedy some of these issues, RashTree is presented as an alternative. Because it is based in newer statistical models and the DIF detection is presented visually, RaschTree is a viable alternative to all current methods.

To summarize the general findings of this study, Rasch Tree could be explored as an additional alternative to detecting DIF than previously proposed methods. The study illustrates how RaschTree can include multiple variables, which many other commonly used DIF detection strategies typically cannot do. In addition, most other DIF detection methods do not allow for the inclusion of continuous predictor variables, which is not a problem for RaschTree, given that it will automatically detect appropriate cut-points in covariates without their being pre-specified. It also demonstrates how RaschTree can detect interactions between variables, which other DIF detection strategies either simply cannot do or cannot do well. Lastly, RaschTree is useful in the sense that it provides a visual way to look at DIF. Other detection strategies do not provide visuals. With RaschTree, every variable included shows just where the exploratory cut-point was found, as well as any interactions between those variables.
The current study provides evidence that RaschTree allows for inclusion of multiple variables at once by using multiple continuous variables. Because RaschTree is based on recursive partitioning, cut points for focal and reference groups do not have to be specified beforehand and allows for a more exploratory approach. Because the recursive partitioning will find cut points based on defined splits, differences in model parameters are able to be partitioned based on stability of the DIF variable. If the parameters were unstable, RaschTree splits the sample for the DIF variable with the greatest instability using model parameters leading to the highest possible stability. Essentially, if the model rejects the null hypothesis (in this case, that there were no differences between groups), there would be only one node in the visual representation of the RaschTree. If there is a split in the groups, that split was found as a result of stabilizing the model based on the largest amount of DIF.

Other DIF detection strategies struggle to allow for inclusion of multiple variables because split points must be predefined. If there is a variable included in the RaschTree, there does not need to be a focal and reference group that is defined beforehand. With other detection strategies, the researchers must have a pre-defined group that they are assessing for DIF (the focal group) as opposed to the group in which DIF is not present (the reference group). Because this is not the case with RaschTree, variables can be included that DIF is suspected but that researchers are not sure which group would be the focal group. Therefore, many different exploratory factors that can be evaluated and detected as having DIF. This is helpful in the sense that a variable that may be negatively affecting one group (for example,
reading enjoyment), can be established and steps to remedy the situation in the classroom can begin to be taken. There have also been shown to be issues with Type I error control if more than one variable is included in the model, with some detection strategies not rejecting DIF models, which makes it difficult to know which items have DIF and which are testing falsely positive. For those detection strategies that do an adequate job of controlling for Type I error rates, there are other issues that RaschTree seeks to remedy, such as having to fit multiple models to the data. Also, while latent class analysis can detect DIF in non-pre-determined groups, they are not directly interpretable.

This study also illustrates how RaschTree can effectively detect interactions between variables, which other models either are not able to do or do not do well. While there is some evidence that certain other DIF detection strategies can include more than one variable, other issues arise when this is the case. In order to detect interactions between variables, more than one variable must be included in the IRT model. Currently, multiple IRT models must be fit to data. With RaschTree, however, not only can more than one variable be included (in this case, three), but interactions between all three variables can be directly interpreted as well as given visual output that includes where the first DIF is detected and where other cut-points are that include the interactions of multiple variables. This is important because presently, only one variable can be tested at a time for DIF. In reality, there are usually multiple variables at play when it comes to testing and assessment. For example, in a classroom that may have children from different home lives, it is possible (and perhaps even likely) that there may be DIF on reading assessment
based on gender but also how engaged they are in reading (based on what their parents have time for at home). Often, there is more than one variable at play for a variety of things, including classroom assessments. This study seeks to provide evidence that RaschTree is able to remedy a gap in DIF detection where interactions are not detected by allowing for groups of items and groups of subjects affected by other dimensions to be identified.

Finally, this study shows just how RaschTree provides illustrated output that other DIF detection strategies do not. RaschTree does provide numerical values for item parameters and where splits in groups are automatically detected, but also provides a tree that illustrates those values visually. With every item with DIF that RaschTree detects, there is a first node with the significance value included. That can grow from one variable (in this case, gender), to two (confidence and enjoyment), to multiple variables (gender, confidence, and engagement), all giving a visual representation of each variable, where the significant split was shown, a significance value of that particular split, down to terminal nodes that illustrate estimates of item difficulty for each item included in the assessment. This is a crucial example of something that makes RaschTree a better option for DIF detection: not only are the results directly interpretable, but the interpretation is made easy to visualize thanks to the ‘tree’ illustration that shows numerical values.

Currently, there is no one strategy that works best for detection of DIF. What test works best is completely dependent on a number of variables, including power, sample size, Type I error rate, number of variables, models tested, and DIF
percentage. RaschTree is presented as a new detection strategy with multiple benefits compared to other proposed methods.
References


the fourth and eighth grades. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.


