THE PERFORMANCE OF THE RASCH TREE METHOD FOR DETECTION OF DIFFERENTIAL ITEM FUNCTIONING IN THE PRESENCE OF MISSING ITEM RESPONSES

By

AHMET GUVEN

A THESIS PRESENTED TO THE GRADUATE SCHOOL OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ARTS IN EDUCATION

UNIVERSITY OF FLORIDA

2017
To my family
ACKNOWLEDGMENTS

When I was an undergrad in the department of Science Education at Ahi Evran University, I felt I was the soul of the darkness, and did have no idea about where I was and had to do. I was trying to find a glint to escape from that feeling. Fortunately, I found myself into questioning the inequality among the others during my course named "Instructional Technology & Material Development." I have learned that the inequality is not only due to individual differences but also lies in whether being an openminded person to change or not. Firstly, I would like to thank Dr. Ahmet Kilinc who encouraged me to move on my way no matter what the others think. Second, I would like to thank the Republic of Turkey which has provided me everything I need. Next, I would like to send my appreciations to Ozgur Erbek who let me know about the study abroad program. Also, I would like to show my grateful Asst. Prof. Dr. Sungur Gurel who encouraged me to keep going without fear when I faced the difficulty at the beginning of my career in America as an unexperienced individual. Lastly, during this demanding time, I would like to thank Ozer Culhaoglu, Fatih Yasar, Nebi Yesilekin, Oktay Keten, Serhat Kara, Buse Tuba Yilmaz, Eyup Civelek, Duygu Berber, Fikret Aydemir, Tolunay Aydemir for supporting me and showing me what I can do.

During my master degree, I would like to thank Dr. Anne Corinne Huggins-Manley who guided me through a priceless master experience so that I can find the solutions by myself when I stumbled into. She taught me the way which I can improve my academic skill and social ability instead of giving the answer right away. Most importantly, I have learned how important to overcome an issue by myself, and how I must be when the time comes to advising my students. I must admit that I would not be able to excel in what I have done without her support.

Also, I would like to present my grateful to Dr. David Miller for being a part of my graduate school experience. Whenever I need to ask something for him, he never hesitated to
answer my questions, and always meet me with a smiling face. When I took his class about the theory of measurement, I realized that he has great knowledge, and is a full of experiences that I made me feel that I must study hard.

Last but not least, I would like to thank Aydin Guven, Ayse GUVEN, Suzan Guven, Kezban Sozer, Mert Guven, and Nehir GUVEN for being a part of my life, and for listening to me at every step of my decisions without getting tired. Since I was born, I have always felt their constructive supports at my heart. I am so happy to be a member of this amazing family, and I must admit that what brought me this point is their love.

Before closing this part, I would like to thank the one who always makes me smile and look at the world with love. She is the one who has always supported me to put myself together and has given me the most amazing glance which fills through my all body. Nihal Elvanoglu, you are one of the best things I have ever had. When you are around, the life becomes better. When you talk, the sound is so kind. When you walk, the distances get closer. I am so glad I have you.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>4</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>8</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>9</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>10</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>12</td>
</tr>
<tr>
<td><strong>CHAPTER</strong></td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>13</td>
</tr>
<tr>
<td>2 THEORETICAL FRAMEWORK</td>
<td>16</td>
</tr>
<tr>
<td>3 METHOD</td>
<td>45</td>
</tr>
<tr>
<td>4 RESULTS</td>
<td>51</td>
</tr>
</tbody>
</table>

**DIFFERENTIAL ITEM FUNCTIONING**

- Differential Item Functioning: 16
- Missing Data: 19
- Treatments for Missing Data: 21
- Detecting DIF in the Presence of Missingness: 23
- Introduction to Tree-Based Methods: 27
- Overview of Item Response Theory: 31
- The Combination of IRT and Tree-Based Models: 33
- Related Studies about Item Response Tree Model: 37
  - Testing DIF Hypotheses at the Item Level: 37
  - Testing DIF Hypotheses at the Test Level: 41
- Research Questions: 43

**SIMULATION METHOD AND SETTINGS**

- Simulation Method and Settings: 45
  - Sample Size and Number of Items: 45
  - Person Parameters: 45
  - Item Parameters: 46
  - Predictor Variables: 46
  - Percentages of Missing Data: 46
  - Type of Missing Data: 47
- Analyses: 48
  - Evaluation Criteria: 49
  - Type 2 Error Rate: 49
  - Type 1 Error Rate: 50
Type 1 Error Rate ........................................................................................................................................51
Type 2 Error Rate or Power ..........................................................................................................................52

5 DISCUSSION .............................................................................................................................................56

APPENDIX

A SIMULATION CODES FOR MCAR MECHANISM .....................................................................................60
B SIMULATION CODES FOR MAR MECHANISM .........................................................................................63
C SIMULATION CODES FOR MNAR1 MECHANISM ..................................................................................65
D SIMULATION CODES FOR MNAR2 MECHANISM ..................................................................................67
LIST OF REFERENCES ................................................................................................................................70
BIOGRAPHICAL SKETCH ............................................................................................................................77
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>The method testing DIF at the item level</td>
<td>38</td>
</tr>
<tr>
<td>2-2</td>
<td>The method testing DIF at the global level</td>
<td>41</td>
</tr>
<tr>
<td>4-1</td>
<td>Type 1 and Type 2 error rates for the Rasch tree method by percentages of missing data, and type of missing data</td>
<td>53</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>2-1</td>
<td>An example of tree structure</td>
<td>29</td>
</tr>
<tr>
<td>2-2</td>
<td>An example of ICC</td>
<td>33</td>
</tr>
<tr>
<td>2-3</td>
<td>Tree-based approach parameterized with IRT model</td>
<td>36</td>
</tr>
<tr>
<td>2-4</td>
<td>An example of the Rasch tree model obtained in the present study with one binary covariate</td>
<td>43</td>
</tr>
<tr>
<td>4-1</td>
<td>Type 1 error rates across all conditions</td>
<td>54</td>
</tr>
<tr>
<td>4-2</td>
<td>Power rates across all conditions</td>
<td>55</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>AAMRT</td>
<td>Auto-Associative Multivariate Regression Trees</td>
<td></td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
<td></td>
</tr>
<tr>
<td>BLRT</td>
<td>Bootstrap Likelihood Ratio Test</td>
<td></td>
</tr>
<tr>
<td>BHM</td>
<td>Begg MH</td>
<td></td>
</tr>
<tr>
<td>CART</td>
<td>Classification and Regression Tree</td>
<td></td>
</tr>
<tr>
<td>CFA</td>
<td>Confirmatory Factor Analysis</td>
<td></td>
</tr>
<tr>
<td>CISIB</td>
<td>Crossing Simultaneous Item Bias Test</td>
<td></td>
</tr>
<tr>
<td>DIF</td>
<td>Differential Item Functioning</td>
<td></td>
</tr>
<tr>
<td>DSF</td>
<td>Differential Step Functioning</td>
<td></td>
</tr>
<tr>
<td>E-M</td>
<td>Expectation-Maximization</td>
<td></td>
</tr>
<tr>
<td>GRM</td>
<td>Graded Response Model</td>
<td></td>
</tr>
<tr>
<td>IFT</td>
<td>Item Focused Tree</td>
<td></td>
</tr>
<tr>
<td>IRT</td>
<td>Item Response Theory</td>
<td></td>
</tr>
<tr>
<td>IRT-C</td>
<td>Item Response Theory with Covariates</td>
<td></td>
</tr>
<tr>
<td>IRT-C</td>
<td>Item Response Theory with Covariates</td>
<td></td>
</tr>
<tr>
<td>ICC</td>
<td>Item Characteristic Curve</td>
<td></td>
</tr>
<tr>
<td>LCA</td>
<td>Latent Class Analysis</td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>Logistic Regression</td>
<td></td>
</tr>
<tr>
<td>LRT</td>
<td>Likelihood Ratio Test</td>
<td></td>
</tr>
<tr>
<td>MAR</td>
<td>Missing at Random</td>
<td></td>
</tr>
<tr>
<td>MCAR</td>
<td>Missing Completely at Random</td>
<td></td>
</tr>
<tr>
<td>METAMH</td>
<td>Meta-Analytic Adjustment</td>
<td></td>
</tr>
<tr>
<td>MH</td>
<td>Mantel-Haenszel</td>
<td></td>
</tr>
<tr>
<td>MI</td>
<td>Multiple Imputation</td>
<td></td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>MNAR</td>
<td>Missing Not at Random</td>
<td></td>
</tr>
<tr>
<td>MPT</td>
<td>Multinominal Process Tree Modeling</td>
<td></td>
</tr>
<tr>
<td>MRT</td>
<td>Multivariate Regression Trees</td>
<td></td>
</tr>
<tr>
<td>NUDIF</td>
<td>Nonuniform Differential Item Functioning</td>
<td></td>
</tr>
<tr>
<td>PCM</td>
<td>Partial Credit Model</td>
<td></td>
</tr>
<tr>
<td>PLRT</td>
<td>Penalized Likelihood Ratio Test</td>
<td></td>
</tr>
<tr>
<td>PMH</td>
<td>Pommerich MH method</td>
<td></td>
</tr>
<tr>
<td>PML</td>
<td>Penalized Maximum Likelihood</td>
<td></td>
</tr>
<tr>
<td>RSM</td>
<td>Rating Scale Model</td>
<td></td>
</tr>
<tr>
<td>SRI</td>
<td>Stochastic Regression Imputation</td>
<td></td>
</tr>
<tr>
<td>SIBTEST</td>
<td>Simultaneous Item Bias Test</td>
<td></td>
</tr>
<tr>
<td>UDIF</td>
<td>Uniform Differential Item Functioning</td>
<td></td>
</tr>
<tr>
<td>1-2-3-PL</td>
<td>One-Two-Three Parameter Logistic</td>
<td></td>
</tr>
</tbody>
</table>
THE PERFORMANCE OF THE RASCH TREE METHOD FOR DETECTION OF DIFFERENTIAL ITEM FUNCTIONING IN THE PRESENCE OF MISSING ITEM RESPONSES

By

Ahmet Guven

August 2017

Chair: Anne Corinne Huggins-Manley
Major: Research and Evaluation Methodology

In this study, the performance of the Rasch tree differential item functioning (DIF) method in the presence of missing data was investigated. We used three types of missingness mechanisms with different proportions of missing cases: Missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). We also created two different MNAR cases to analyze their differential effects on the DIF detection. We conducted a simulation study manipulating percentages of missing data and type of missing cases, along with a single item containing a large amount of DIF. The results indicated that the Rasch tree method for the detection of DIF had high power under all types of missing cases and all percentages of missing data, even if the method removed the nonresponses in a target item with listwise deletion. However, under one of the MNAR conditions with at least 5% of missing cases, the Type 1 error rates were elevated whereas under other MNAR condition with 15% of missing cases, the error rate was highest in all conditions. In addition to that, under MCAR and the MAR conditions, the power rates slightly declined as the proportions of missing cases increased from 5% to 30%.
CHAPTER 1
INTRODUCTION

Missing item responses have been a common problem in statistical analysis, which can have a substantial impact on statistical conclusions. For instance, respondents might not provide a response on a scale item for a systematic reason such as fatigue, or for a random reason. Another issue in measurement data can be the presence of differential item functioning (DIF), which occurs when the conditional measurement properties of a test differ across groups of respondents. Both concerns have been investigated separately in the educational literature (e.g., Tutz & Schaubberger, 2015; Tay, Newman, & Vermunt, 2011; Lee, 2017; Swaminathan, & Rogers, 1990; Pohl, Grafe, & Rose, 2014; Finch, 2008; Kohler, Pohl, & Carstensen, 2014; Vidotto, Vermunt, & Kaptein, 2015) and have also been studied in conjunction (e.g., [provide a couple of citations from studies that examined both DIF and missingness at the same time]). If there is a statistically significant difference in conditional item characteristics, such as item difficulty, across groups of respondents, one has access to many methods to detect such a problem of DIF to identify and prevent issues that may lead to item bias. The plethora of available methods for DIF detection is summarized in several books (e.g., Holland & Wainer, 1993; Osterlind & Everson, 2009), chapters (e.g., Penfield & Camilli, 2007), and manuscripts (e.g., Clauser & Mazor, 1998). The literature on DIF continues to grow, as evidenced by many recent research studies. For instance, Lee (2017) examined the performance of the logistic regression (LR) DIF method in small samples by using different statistical inference approaches. Also, Tutz and Schaubberger (2015) studied the DIF detection of a penalty approach in the Rasch model, which allows for the testing of DIF across groups defined by multiple variables. Directly related to the present study, several DIF methods based on recursive partitioning models have been proposed by Strobl, Kopf, and Zeileis (2015), Jeon and De Boeck (2016), Tutz and Berger
(2016), and Berger and Tutz (2016). This line of DIF methodology research proposes several different tree-based, data mining methods for evaluating parameter invariance across groups of subjects. To date, these model-based recursive partitioning approaches have shown high power for the detection of DIF.

As per missing item responses in measurement data, many treatments have been investigated to check the impact of missing methods on parameter estimation quality. Varied methods for dealing with missing measurement data have been proposed to see how well they treat missingness in the context of different missing mechanisms with different percentages of missing data (as well as other test and data factors). Finch (2008) used several different missing treatment methods to compare the resultant bias in item discrimination, difficulty, and guessing parameters from item response theory (IRT) models under various conditions. The study demonstrated that the degree of parameter estimation bias depends not only on the method used to treat missing data but also the type of missing mechanism. Especially when data were missing not at random (MNAR; i.e., the missingness is systematically related to the trait being measured), greater bias in parameter estimation is expected but also varies depending on the method used for treating missing data.

It is a reasonable assumption that social science researchers and practitioners may collect item response data that consists of both item nonresponse (i.e., missing data) and some presence of DIF. In this context, it is critical that such researchers and practitioners have access to research that studied DIF detection methods in the presence of missingness. However, only a few studies have considered investigating the detection of DIF when missingness is present (e.g., Finch, 2011a, 2011b; Robitzsch & Rupp, 2009; Emenoglu, Falenchuk & Childs, 2010). They compared different types of DIF detection methods with various missing treatments under various
missingness mechanisms, along with some other testing and data factors such as sample size. The general question was about how much the performance of DIF detection methods is affected by the presence of missing data. In the present study, this line of research is extended to tree-based DIF detection methods. Specifically, we explore the impact of missing data on the performance of the Rasch tree DIF detection method proposed by Strobl et al. (2015).
CHAPTER 2
THEORETICAL FRAMEWORK

Differential Item Functioning

A well-known phenomenon in psychometrics and educational measurement is differential item functioning (DIF), which is a major issue that can lead to biased measurement. DIF is defined as a difference in conditional probabilities of response on a given test item based on some external grouping variable (Holland & Wainer, 1993). Two groups are often used when conducting DIF analyses, which are referred to as the reference and focal group. If DIF is present in a binary test item, then one of the groups has a disadvantage on the item. Theoretically, DIF occurs when items on an assessment tool or psychological test are multidimensional. That is, in addition to the primary dimension measured by the tool there must be some additional dimension that is being measured, one on which the groups have differential distributions (Penfield & Camilli, 2007).

Two general types of DIF effects have been studied in recent decades: uniform and nonuniform DIF. The DIF effect is said to be uniform when the direction (i.e., which group is favored) and magnitude of the effect is consistent across the underlying trait range. On the other hand, if the differential conditional probabilities across groups vary in direction and/or magnitude across the trait range, then nonuniform DIF effect is present (Penfield & Camilli, 2007). While the primary purpose of conducting DIF analyses is to capture the items showing differences in their underlying ability distributions, several studies indicate that it can be seen as one step in the larger investigation of construct validity (Walker, Beretvas & Ackerman, 2001; Walker, 2011; Walker & Gocer Sahin, 2016).

Many methods of DIF detection have been developed and studied, such as the Mantel-Haenszel (MH) statistic developed by Mantel and Haenszel (1959), the penalty approach
proposed by Tutz and Schaubbeerger (2015), standardization procedures (Doran & Kullick, 1983), logistic regression (LR) methods (Swaminathan & Rogers, 1990), likelihood ratio test method (LRT; Thissen, Steinberg & Gerrard, 1986), the MIMIC model method (Finch & French, 2011), and more. All the methods can be classified into one of two groups: Non-IRT methods and IRT-based methods. The distinction between methods lies in whether they are based on IRT (and a more in-depth review of IRT is covered in a later section of this chapter). It has been discussed that the two types of methods can show discrepant results from a DIF analysis (Stark, Chernyshenko & Drasgow, 2006). There are some statistical and rational arguments for choosing between these two types of methods, but a discussion of this is beyond the scope of the current study. For this literature review, some studies that introduced and/or compared various DIF methods are summarized here such that the Rasch Tree method examined in this study can be understood in relation to some other existing methods and DIF studies.

Swaminathan and Rogers (1993) examined the performance of the MH and LR methods on the detection of nonuniform and uniform DIF. They articulated that the LR method can be thought of as a logistic regression model with the discrete ability variable, a grouping variable, and an interaction term. They found that the logistic regression procedure outperformed over the Mantel-Haenszel on the detection of nonuniform DIF. However, the two methods performed similarly on the detection uniform DIF.

Westers and Kelderman (1992) proposed a method for the detection of DIF in multiple choice tests to avoid some issues with other DIF methods that do not control for ability with an IRT model. The proposed model, which was formulated by a latent class analysis (LCA), distinguishes between a "Know" state and a "Don`t Know" state. The "Know" state is in which respondents completely know the answer while the "Don`t Know" state indicate the respondents
will select the most attractive alternatives. They demonstrated that the main advantages of this method was not only to test an individual item for DIF but also to test the impact of item difficulty on DIF.

French and Finch (2013) examined uniform DIF in the presence of multilevel data with variations to the MH procedure. These new methods, such as the Begg MH approach (BMH: Begg, 1999) and its variations based on the magnitude of the ratio of variances, the Pommerich MH method (PMH), and the MH based on the meta-analytic adjustment (METAMH), were compared with the traditional MH method using a simulation study in the case where the multilevel data structure occurs, such as the case in wherein the examinees nested in schools. They stated that the MH procedure has some advantages and few disadvantages in practice such as it outperforms the methods based on IRT and logistic regression when the sample size is not large enough. Also, the proposed method can be conducted with common statistical analysis programs.

Stark, Chernyshenko, and Drasgow’s (2006) study focused on the comparison of two different methods based on the LRT DIF test, which can be utilized within both CFA and IRT. They examined the loadings and the intercept parameters simultaneously using LRT and Bonferroni corrected critical p values. They indicated that the proposed method was stronger than an alternative approach in the case where DIF was implemented on item thresholds.

Lee (2017) studied on determining the extent to which the logistic regression procedure (LR) relying on the maximum likelihood estimation and an asymptotic assumption is reliable in small samples. He also examined that other methods could be considered as alternatives such as the penalized maximum likelihood (PML) estimation, penalized likelihood ratio test (PLRT), and
bootstrap likelihood ratio test (BLRT). The simulation study showed that the LRT based on the asymptotic chi-square distribution is powerful even in a small sample size.

Tay, Huang, and Vermunt (2016) investigated the extent to which the IRT with covariates (IRT-C) model (Tay, Newman & Vermunt, 2011) can be used to detect DIF across multiple covariates in the environment where the test was implemented to large sample sizes. They found that the procedure showed the power to detect DIF across all covariates even in large sample sizes. Also, it worked well on the recovery of three-parameter logistic model parameters and the latent means.

In order to investigate further issues with respect to DIF, Cho, Suh and Lee (2016) pointed out what happens after the detection of DIF. They evaluated confirmatory multigroup multidimensional item response model to provide the rationale about the accuracy of impact estimates, parameter estimates, and person scores on the leading dimension while controlling for the second-dimension due to DIF. Moreover, the item response model approach was compared with the practices such as deleting and ignoring DIF items. Results indicated that modeling DIF approaches were more effective than the others for recovery of item parameters and person scores. In addition, they reported that the modeling DIF approach could be a reasonable method to treat DIF items.

**Missing Data**

The problem of the existence of missing data in responses on measurement tools has been discussed extensively in the literature, and it is generally known that missing data makes the statistical analysis of a given data set much more complicated as compared to the case of complete data (Glass & Pimental, 2008; Finch, 2008; Holman & Glass, 2005). Therefore, there is an ongoing interest in studying methods that handle missingness and account for this incompleteness. What makes the respondents not respond to items has been studied in a thorough
way such as the case where missing data is due to random and ineffective designs (Glass & Pimental, 2008; Finch, 2008; Holman & Glass, 2005; Moustaki & Knott, 2000; Glass, Pimental & Lamers, 2015; Rose, Von Davier & Nagengast, 2015). According to Little and Rubin (2014),

three types of missing data mechanisms can be seen in a data set, including missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR refers to the case where the missingness is not associated with any systematic mechanism and occurs in an entirely random fashion. According to Rose, Von Davier and Nagengast (2015), the mechanism is MCAR if

\[ D \perp (Y, I, \xi) \]  \hspace{1cm} (2-1)

Let \( Y \) be a random matrix containing binary random variables \( Y_{ni} \) that codes the item response for the \( n \)th person and \( i \)th item. \( D \) indicates the observational status of \( Y_{ni} \), and \( \xi \) is the latent ability. \( I \) refer to covariates in the model. Hence, Equation 2-1 states that the presence of missing data on an item is independent of the latent ability, the covariates in the model, and item responses.

MAR points out that the probability of an observation including non-response is related to an observed variable. For example, in some cases, male respondents are more likely to leave a particular item unanswered than are females. As provided by Rose, Von Davier and Nagengast (2015), If the data is called to MAR, then with covariates \( I \),

\[ D \perp (Y, \xi) \mid I. \]  \hspace{1cm} (2-2)

Equation 2-2 states that the missing data in an item is related to a covariate, but after conditioning on that covariate, the missing data is unrelated to the item items and the underlying latent trait.
Lastly, MNAR refers to the case where the probability of missingness is associated with the variable itself. For example, if an examinee does not respond to an item because of a lack of ability to answer the item (i.e., the same ability that is being measured by the test), data is called to MNAR. The MNAR refers to the following definition (Rose, Von Davier & Nagengast, 2015);

\[
D_{\text{MNR}} \left( Y_m, \xi \right) \mid \left( Y_o, I \right)
\]

Let \( Y_m \) and \( Y_o \) be an observed part and the missing part, respectively, which are obtained by decomposing the observed data matrix \( Y=y \). In this case, MNAR is considered as a nonignorable type of the missing data mechanism since it might lead to biased estimates of the parameters. Omitted and not reached items might usually be seen as nonignorable. In order to avoid biased estimates, the missing responses need to be handled in some methodological manner. Several studies have been conducted to distinguish the performance of different methodological approaches to missingness under various types of missing mechanisms and other testing/data conditions (Culbertson, 2011; Ayala, Plake & Impara, 2001; Finch 2008; Holman & Glas, 2005; Glas & Pimental, 2008). In the following section, some selected methods of dealing with the missing data are briefly summarized in minor detail, as further details is beyond the scope of this study.

**Treatments for Missing Data**

There are three general types of approaches to dealing with the missing item response data: classical approaches, imputation-based approaches, and model-based approaches.

Traditional approaches deal with the missing values by either deleting them from the data in various ways (including listwise deletion, pairwise deletion, and more) or replacing the missing data with incorrect scores (assuming the test is an achievement test). Deleting missing data makes a stringent assumption that data are MAR or MCAR. Replacing the data with incorrect
scores assumes that the respondents do not know the answer. This approach has a significant drawback that might lead to biased parameter estimates (Newman, 2003; Lieberman-Betz et al., 2014; Acock, 2005; Finch, 2011a, 2011b). For this reason, Lord (1974) proposed that missing responses might be scored as fractional correct. Many studies have been conducted to investigate the performance of several types of traditional approaches. Only scoring the values as fractional correct gives almost accurate results as a complete dataset exists. However, when the ignorability does not hold, the approaches should be picked carefully. Many studies have pointed out that even if the data are MCAR, classical approaches such as listwise deletion might result in biased estimates (Enders 2010; Lieberman-Betz et al., 2014; Finch, 2011a, 2011b).

In imputation-based approaches, missing responses are replaced by valid response options based on some method for determining the most likely valid response the person would have provided. These methods include corrected mean substitution (Bernaards & Sijtsma, 2000; Huisman & Molenaar, 2001; Sijtsma & Van der Ark, 2003), response function imputation (Sijtsma & Van der Ark, 2003), imputing missing data with expectation-maximization algorithm (E-M algorithm, Dempster, Laird, & Rubin, 1977), and multiple imputation (MI; Rubin, 1996). According to Finch (2008) ’s study, MI, which is a Bayesian approach where the values are drawn from the posterior distribution, outperforms all other imputation methods. However, when the IRT models are used MI has a substantial disadvantage since it requires repeated parameter estimation (for further details, see Finch (2008)).

As for model-based approaches, two different approaches have been recently studied to account for the nonresponse mechanism: latent approach (O’Muircheartaigh &Moustaki, 1999; Glas & Pimentel, 2008; Moustaki & Knott, 2000; Holman & Glas, 2005) and manifest approach (Rose, von Davier, & Xu, 2010). According to these studies, missing propensity and ability are
correlated to each other. That is, missing responses due to omission and the speediness of the test is associated with the person ability, and missing mechanism might not be ignorable. In latent approach, not-reached and omitted items are modeled in a different way while the manifest approach can be used for both types of missing responses. The model adjusted by Glass and Pimental (2008) based on Holman and Glass (2005)’s model can be used to model missing responses due to not-reached items. In the manifest approach, the latent approaches might cause an estimation problem especially when the sample size is small and when the multidimensionality between the latent ability and the latent response tendency exists (Pohl, Grafe & Rose, 2014; Rose, Von Davier & Nagengsat, 2015).

**Detecting DIF in the Presence of Missingness**

The purpose of this study is to examine the performance of the Rasch tree method for DIF detection when missing data is present. In the test environment, missing responses appear due to several reasons such as skipping items after reading in the given time limit. This missingness can pose issues for the investigation of measurement invariance when DIF analyses are conducted between or across the groups who tend to omit or skip the items. Thus, there is a need to understand how well a DIF method performs in the presence of missing data. Since DIF methods such as MH and LR suffer from the existence of missing data (Banks, 2015), choosing a particular treatment for missing cases becomes an important matter in the context of flagging potential item bias. For instance, removing (listwise) the respondents who have some missing item responses could cause a remarkable reduction in the sample size and increase the Type I error rate if no DIF item exists (Finch, 2011a, 2011b). A few studies have been conducted to examine the impact of missing data on the detection of DIF (Banks, 2015; Robitzsch & Rupp, 2009; Finch, 2011a, 2011b; Emenoglu, Falenchuk & Childs, 2010). Data generation and
manipulation conditions of the studies in DIF context in the presence of missing data will be given at the end of each paragraph.

Emenogu, Falenchuk, and Childs (2010)’s study investigated the effect of the method of handling missing data when the differential nonresponse rates exist after conducting the M-H DIF analyses. According to the conditions given below, the study suggested that if there was little loss in the sample size, the choice of missingness method would not be a problem. On the other hand, if missingness caused a great reduction in the sample size, differential nonresponse rates might have an impact on the source of DIF. For the simulated condition, changing nonresponse as an incorrect response was not a good option to detect DIF. As per listwise deletion conditions, they found that there were no flagged items exhibiting DIF since the observations having nonresponses reduced the power of the analyses of detection DIF. The result of the study suggested that using analysiswise deletion could be the best option for detecting DIF in the presence differential nonresponse across groups. The conditions in the study were in the following: 1) The number of item and sample size were 25 and 2000 for each group, respectively. 2) The proportion of missing data was not explicitly given. 3) Missing data treatments were listwise deletion, zero imputation (ZI), and analysiswise deletion (AD). 4) Missing mechanism was MNAR. 5) DIF method was M-H. 6) There was no given information about magnitude of DIF.

Robitzsch and Rupp (2009) investigated how zero imputation, listwise deletion, response function imputation, and two-way imputation for handling with missing item responses interact with two DIF methods (the M-H and the LR). The effect of the choice of DIF detection methods was not investigated. Three types of missingness were generated in the simulation studies to see to what degree they would have an impact on Type I error rates and Type 2 error rates. The
results showed that the choice of missing treatment methods could have a bearing on the bias and the Type 1 error rates either when no DIF is present or when DIF exists. Moreover, they found that missing data treatments under the conditions where data are MCAR seemed less problematic than the conditions in which MNAR was the case. They pointed out that zero imputation method can influence the results, which means that it can lead to biases if the data set includes plenty of nonzero responses. The following statements explained the conditions related to the present study: 1) The number of item and sample size were 20, and 40 and 250, 1000, and 4000 for each group, respectively. 2) The proportion of missing data was determined as 0%, 10%, 30%, and 30%. 3) Missing data treatments were multivariate imputation by chained equations (MICE), listwise deletion, ZI, Two-Way, and Two-Way Adjusted. 4) Missing mechanisms were MCAR, MAR (two types), and MNAR. 5) DIF method was LR, and M-H. 6) Difficulty parameter was used to determine magnitude of DIF (0.2 as small, 0.4 as medium, and 0.6 as large).

Finch (2011a) used three methods to detect DIF items when multiple imputation methods are utilized for handling missing dichotomous data. Below will explain the simulation conditions and the methods for both the detection of DIF and missing data. So as to determine how the manipulated factors, or interactions among these factors, influenced the power and Type I error rates for three DIF methods, ANOVA models were conducted. The study suggested that when data were MAR, considering missing data as an incorrect response severely inflated the Type 1 error rate. However, the listwise deletion and MI (multiple imputation) methods for handling missingness were associated with Type I error rates of DIF that were at the nominal level of 0.05. They suggested that with respect to the missing method, the listwise deletion approach might be preferred to MI in several aspects since the listwise deletion is the most accessible method in many software if the primary goal is to detect DIF items correctly. In terms of DIF
detection methods, only in the smallest sample size and the case where there is %15 missingness, the LR had a slightly high power as compared to the others. On the other hand, in other conditions, DIF methods showed a close performance to one another. The conditions in the study were in the following: 1) The number of item and sample size were 40 and 250, 500, and 1000 for each group, respectively. 2) The proportion of missing data was determined as 0%, 5%, and 15%. 3) Missing data treatments were listwise deletion, ZI, and MI. 4) Missing mechanisms were MCAR, MAR, and MNAR. 5) DIF method was M-H, simultaneous item bias test (SIBTEST), and LR. 6) Item difficulty was used to determine magnitude of DIF (0.3 as medium, and 0.6 as large).

In contrast to Finch (2011a), Finch (2011b) examined the performance of three methods on the detection of nonuniform DIF when missing item responses are present in various degrees and types (see below). ANOVA was used to determine if the manipulated factors in this study were associated with power and Type 1 error rates for the DIF methods. Four types of missing mechanisms were used to investigate to what degree DIF detection methods are influenced if the missing data varies in the percentages of the amount. The power results of the ANOVA indicated that the five-way interaction including missing data methods, amount of missing data, a level of DIF, impact, and missing data mechanism had the highest order significance in the model. ANOVA for the Type I error rates demonstrated that the three-way interaction containing missing data method, percentages, and missing data mechanism showed the highest significant result. Furthermore, the case in which data were MAR 2, which means that a likelihood of leaving an item blank might be greater in a group of respondents, had more powerful when listwise deletion was used to deal with missing data. In general, the result implied that using the listwise deletion method for the treatment of missing data was as accurate as in the complete data.
set regarding Type 1 error rates, effect size, and power. The following statements explained the conditions related to the present study: 1) The number of item and sample size were 20, and 40 and 250, 500, and 1000 for each group, respectively. 2) The proportion of missing data was determined as 0%, 10%, 20%, and 30%. 3) Missing data treatments were stochastic regression imputation (SRI), ZI, and listwise deletion, and MI. 4) Missing mechanisms were MCAR, MAR (two types), and MNAR. 5) DIF method was = item response theory likelihood ratio (IRTLR), crossing simultaneous item bias test (CSIB), and LR. 6) Discrimination was used to determine magnitude of DIF (0.4 as small, 0.8 as medium, and 1 as large).

As will be shown in a minute, new tree-based DIF methods have been proposed and studied in recent psychometric literature. However, no study to date has tested the performance of these DIF methods in the presence of missing data. Just as the studies above that examined the performance of non-tree based DIF methods in the presence of missingness were critical for determining the extent to which such DIF methods can be used by practitioners who have missing data, the tree-based DIF methods are in need of such research. First, an introduction to tree-based methods, IRT modeling, and IRT-trees is needed.

**Introduction to Tree-Based Methods**

This study will focus on the use of one of the IRTree methods for DIF detection when missing data is present, and hence an introduction to tree-based methods is needed. A well-known algorithm that can be applied to supervised and unsupervised learning environments is known as trees or tree algorithm. A tree algorithm is one of the statistical learning methods whose goal is to create a model that makes a prediction about a test outcome based on several input variables. A tree can be developed by splitting each node on the tree into two daughter nodes. Since tree-based methods partition the predictor space into a set of regions, and require the splitting rules to segment the feature space, these approaches specifically are known as
decision tree methods, which can be applied to both regression and classification problems (James, Witten, Hastie, & Tibshirani, 2013; Hastie, Tibshirani, & Friedman, 2009; Strobl, Malley, & Tutz, 2009; Delgado-Gomez et al., 2016). The main purpose of the tree-based method is to test possible splits and find the regions that minimize the residual sum of squares (RSS) (see Figure 2-1). Formally (James, Witten, Hastie, & Tibshirani, 2013),

\[ \sum_{j=1}^{J} \sum (y_i - y_{rj})^2 \]  

(2-4)

where \( y_{rj} \) is the mean response for the training observation which is used to train the method, \( J \) is the number of the regions after splitting the covariate space, and \( i \) is the observation.

The most common tree-based methods, such as decision trees, differ in their nature with respect to the scale of the outcome variable. If a tree-based method is based on only regression problems, it is called a regression tree. If a tree-based method deals with only classification problems or both regression and classification problems, it is called a classification tree or a regression and classification tree (CART), respectively. These differentiations of types of trees are related to whether one is dealing with a continuous outcome variable or a qualitative outcome variable (James, Witten, Hastie, & Tibshirani, 2013). For example, let’s say that we would like to predict the annual salary in a particular job (i.e., continuous outcome) by using the number of years of experience in that job and the number of overtime hours. In keeping with the tree structure, the years of experience and number of overtime hours variables will be our regions, called \( R_1 \) and \( R_2 \), respectively. They are also known as the terminal nodes or leaves of the tree. A tree structure technically looks like an upside-down tree. That is, the leaves or terminal nodes are at the bottom of the tree. Where the predictor space is split is known as the internal nodes. The branches of the tree provide a connection between the terminal nodes and internal nodes or
between two internal nodes. The starting point of the tree is called the root node (which is also an internal node). Figure 2-1 shows a simple regression tree structure.

![An Example of Regression Tree](image)

Figure 2-1. An example of tree structure

In the example, as we see, the regions or targeted points are $R_1 = \{X|\text{Years} < 5\}$, $R_2 = \{X|\text{Years} \geq 5, \text{Hours} < 4\}$, and $R_3 = \{X|\text{Years} \geq 5, \text{Hours} \geq 4\}$ (e.g., James, Witten, Hastie, & Tibshirani, 2013). Figure 2-1 demonstrates the regions as a function of Years and Hours. Also, years < 5 did not split into more nodes since the tree did not find any significant differences among the observations which belong to years < 5. This example is a regression tree because the outcome variable is continuous. However, the decision trees can be applied to classification problems in a similar way when one might want to deal with a categorical outcome. One possible advantage of tree structures over other types of regression models may be easier interpretation of results for layman audiences.

A set of approaches such as regression tree mentioned above used to estimate $f$ (which refers to the information about the outcome variable over the independent variable) is known as a statistical learning process mentioned above. Either prediction and inference might be a reason for estimating $f$. However, because it is computationally burdening to take account every
possible partition of the covariate space into many boxes, one greedy approach, which is known as recursive binary splitting or recursive partitioning algorithm might be thought of as an alternative to the decision trees (James, Witten, Hastie, & Tibshirani, 2013). This approach might be also called a tree based approach. The RSS equation of recursive partitioning algorithm by James, Witten, Hastie, and Tibshirani (2013) is given in the following:

$$
\sum_{x_i \in R_1(j, s)} (y_i - \hat{y}_{R_1})^2 + \sum_{x_i \in R_2(j, s)} (y_i - \hat{y}_{R_2})^2
$$

(2-5)

where we need the value of the region $j$ and the cutpoint $s$ that minimize the equation. Note that this equation is for regression problems.

Recursive partitioning algorithm based on tree-based models, one of those approaches for estimating $f$, come into play when the tree starts to be built up. They are used to grow a large tree on the training data which is the observations used to understand how the applied method estimates $f$, and then stop growing if there is a stopping criterion provided before applying the algorithm (e.g., James, Witten, Hastie, & Tibshirani, 2013). After that, one wants to apply a pruning mechanism to obtain a sequence of best subtrees. A problem in tree-based models is the possible presence of bias in variable selection (e.g., Loh & Shih, 1997; Hothorn, Hornik, & Zeileis, 2006). Even if pruning successfully selects the size of the tree, the interpretation of the trees can be affected by covariate selection, just as is true in any regression-based application. Omitted covariate misspecifications can ultimately lead to bias in resultant parameter estimates (Hothorn, Hornik, & Zeileis, 2006; Zeileis, Hothorn, Hornik, 2008). Therefore, correct specification of the covariates is a process that requires careful considerations when using tree-based models (Strobl, Boulesteix, & Augustin, 2007). Hothorn, Hornik, & Zeileis (2006) proposed a new framework for recursive partitioning algorithms, one in a theory of conditional inference procedures is used to obtain unbiased variable selection. In the present study, the
method for the detection of DIF is based on this proposed framework that is suggested to
overcome the aforementioned problem.

**Overview of Item Response Theory**

Ultimately, this study will focus on a particular family of regression trees in which IRT
models are embedded. But first, an introduction to IRT is needed. IRT, often called either latent
trait theory or item characteristic curve theory, is a family of statistical models that can serve as
useful tools for analyzing item data to estimate item properties and respondent properties in
psychological and behavioral test environments (Lord & Novick, 1968). IRT is considered a
latent trait theory because the underlying behavior under a mental construct is called latent trait
of the test takers, and person ability ($\theta_p$) reflects the point where an examinee stands on the latent
continuum. IRT models allow us to design and calibrate test items using model based estimates
for person ability and item characteristics, which can later on lead to decision-making process
such as whether or not removing an item that has DIF is appropriate (e.g., Cho, Suh, & Lee,
2016). That is, two informative features of psychological measurement (person ability and item
properties) corresponding to item response patterns can be estimated with IRT models, and the
probability of falling into one particular category can be calculated with the IRT models.

In many but not all cases, IRT models are specified as conventional, multivariate logistic
models (e.g., Baldwin, 2006). The present study will be limited to these specifications. There are
various the IRT models used to develop and analyze the tests and questionnaires. IRT models
were originally developed for the purpose of ability assessment. They can be categorized in two
different ways with respect to the extensions of item formats: binary IRT models and polytomous
IRT models. Each of these model types can further break into two subcategories: unidimensional
models, dealing with only one latent trait, and multidimensional models, where the test measures several latent traits.

Some commonly applied binary IRT models are the Rasch model (Rasch, 1960), two parameter logistic model (Birnbaum, 1968), and three parameter logistic model (Birnbaum, 1968) (e.g., Yan, Lewis, & Stocking, 2004; De Boeck, 2008; Bolt, Deng, & Lee, 2014). The Rasch model, or one parameter logistic model, includes one estimated item parameter (item difficulty, denoted by $b_i$), which represents the location of the item along the latent trait continuum at which the probability of scoring a 1 (as opposed to a 0) for person $p$ with a latent ability ($\theta_p$) is at $P = 0.50$. The probability of obtaining a 1 on an item is conditional on person ability and can be denoted as $P(X_{ip}=1|\theta_p)$. This binary model assumes monotonicity and, therefore, plotting $P(X_{ip}=1|\theta_p)$ will result in an S-shaped curve, which is called Item Characteristic Curve (ICC).

For the two-parameter logistic model, two item properties are freely estimated across items: item difficulty and item discrimination (denoted by the Greek symbol “$\alpha_i$”). Item discrimination indicates the ability of an item to differentiate subjects of different ability levels. For the three-parameter logistic IRT model, a third item parameter is freely estimated, that of a lower asymptote. This parameter is often referred to as a guessing parameter (denoted by the Greek symbol “$\gamma_i$”), as it indicates the degree to which an item as a non-zero lower boundary for its ICC, which would be expected if guessing was occurring on the test items. Figure 2-2 shows an example ICC for one item on a particular test that was modeled under a three-parameter logistic model. The location along the x-axis indicates the item difficulty ($\beta_i = 0.5$), the slope of the S-shaped curve indicates the item discrimination ($\alpha_i = 1$), and the lower asymptote indicates the item guessing ($\gamma_i = 0.3$).
In more recent years, psychometricians are combining tree-based methods and IRT models, such that the parametric model that defines the movement from one node to the next in the tree is an IRT model rather than a linear regression model (Jeon & De Boeck, 2016; De Boeck & Partchev, 2012; Bockenholt, 2012).

The model stated above can be used for various reasons and combinations with the other models to solve the existing problems in the psychological and educational test environment (e.g., Jeon & De Boeck, 2016; Jeon, De Boeck & van der Linden, 2017; DiTrapani, Jeon, De Boeck, & Partchev, 2016; Debeer, Janssen, & De Boeck, 2014; Tutz & Berger, 2016; El-Komboz, Zeileis, & Strobl, 2014; Strobl, Kopf, & Zeileis, 2015; Berger & Tutz, 2016; Bollman, Berger, & Tutz, 2016; Vaughn & Wang, 2010). However, recall from above that there are two general types of decision tree models: classification trees (with the categorical outcome) and regression trees (with the numerical outcome) (James, Witten, Hastie, & Tibshirani, 2013).
difference between two classifications might not be defined with the parenthesized words since logistic regression is applied to a problem with a qualitative response; therefore, it might be thought of a classification problem. However, it still has a regression problem because it deals with the estimation of conditional probabilities through the logistic form of an IRT model. Also, there are some other data mining methods dealing with both qualitative and quantitative outcomes such as K-nearest neighbors and boosting. When it comes to making an inference (e.g. which items contribute to DIF?) and a prediction (e.g. estimating person scores or item difficulties), model-based tree structures, also called recursive partitioning (James, Witten, Hastie, & Tibshirani, 2013), plays an important role in providing information about the data set. Three interwoven versions are known as classification and regression Trees (CART), developed by Breiman, Friedman, Olshen, & Stone (1984), multivariate regression trees (MRT), proposed by De`Ath (2002), and auto-associative multivariate regression trees (AAMRT), proposed by Questier et al. (2005). Additionally, Hothorn, Hornik, and Zeileis (2006) recommended another kind of tree-based models, which is based on conditional inference. They all have been used for several reasons such as estimation accuracy, detecting differential item functioning (DIF), understanding item parameters, and variable selection.

Batchelder and Riefer (1999) described a statistical model for categorical data, which is called the multinominal process tree model (MPT). They explained that the MPT might seem similar to the simplest IRT model regarding parameter estimation, that is, the Rasch model or one parameter logistic model. However, they stated that the MPT and the Rasch model do not serve the same purpose. That is, the Rasch model does not attempt to explain the sequential cognitive processes that respondents may use to respond to test items, whereas that is a core purpose of MPT. An extension of the MPT for discrete choices in psychological measurement
was discussed by Batchelder, Hu, and Smith (2009). Also, a two parameter MPT model was proposed by Schweichert (1993). MPT and its extensions have been widely used in many areas of psychology. One might state that these models are related to a clustering approach to statistical modeling, or one might point out that classification and regression problems are the core basis of these models.

It is evident that tree-based models have been used for measurement research and application for several purposes. One important feature of tree-based models is that they can be utilized for both unsupervised learning (boosting, bagging, clustering e.g.), which does not include outcome variables, and supervised learning where the response variables are available. Moreover, it provides data visualization with the tree structure to help the interpretation of response structure for test items. With the informative characteristic of the decision process in tree-based models, it is likely to note that the IRT models might seem more informative for handling some expected problems in test environments. The advantages of tree structures in measurement modeling may include: 1) dealing with various response types; 2) easiness to interpret measurement invariance due to the nature of tree structure; 3) more informative interpretability of parameters and response processes; and 4) being able to directly handle omitted items or missing data (De Boeck & Partchev, 2012; Jeon & De Boeck, 2015).

Let’s take the example of using IRTrees for directly modeling missing data within an item that otherwise has binary responses (incorrect and correct). Suppose that there is an upside-down tree which includes one root (top node), one following internal node or end node (leading to one or more response categories with each branch), and its branch coming through the internal node at the top (response category corresponding to the latent trait). The probability of going through the branches can be formulated with a logit link or probit link (and as mentioned before,
this study will focus on a logit link). Figure 2.3 shows a linear tree structure or two-stage process tree structure, indicating that $Y_1^*$ node (root node) has one branch (response category “missing”) and one follow-up internal node ($Y_2^*$ for a correct response and incorrect response). Recall from IRT models that the probabilities of responding in particular categories are based on item and person properties. This is logically represented in this IRTree example. Specifically, the tree in Figure 2.3 exhibits two different latent traits: Suppose that there is an achievement test, so $Y_1^*$ can be related to the propensity of omitting the item, and $Y_2^*$ can be related to the ability measured by the test. The probability of choosing a right or left branch from each node is modelled by the following formulas: probit$_{pi}$ or logit$_{pi} = \theta_{b1} + \beta_{i1}$ for $Y_1^*$, and $\theta_{b2} + \beta_{i2}$ for $Y_2^*$. The important point here is that whether $Y_2^*$ is observed depends on the decision at the top node. For the further information about the model formulation, one might want to see the article published by De Boeck and Partchev (2012).

Figure 2-3. Tree-based approach parameterized with IRT model
Related Studies about Item Response Tree Model

In order to provide an explicit explanation about studies that have used item response tree models to evaluate DIF, I have split this section into two parts. The first section summarizes research that uses item response trees to test hypotheses about DIF at the item level (i.e., the null hypothesis is that a particular item does not have DIF, which is a typical DIF null hypothesis in the broader DIF literature). The second section summarizes research that uses item response trees to test hypotheses about DIF at the test level (i.e., the null hypothesis is that all item parameters across the test are invariant across groups).

Testing DIF Hypotheses at the Item Level

De Boeck and Partchev (2012) proposed the models consisting of combinations of item response models and tree-based structures. The models can be subcategorized into two different models: response tree models based on a tree representation of response categories and latent tree models focused on a tree representation of latent variables. Moreover, they stated that IRT models for categorical item responses, where response categories are either dichotomous or polychotomous, and the response tree models, which include a sequential process (moving from the root of a three to its end) proposed by Tutz (1990), can be seen together as a binary response tree. According to De Boeck and Partchev (2012), the intersection at which two models meet falls to the point where a tree based structured models and IRT models can be modeled with either a logit link or probit link. Nevertheless, the tree structures allow for many interesting applications in IRT, such as modeling relationships between fast and slow responses by using a one parameter IRTree model. Table 2-1 shows studies that are related to using tree-based models based on IRT models to detect DIF at the item level.
Table 2-1. The method testing DIF at the item level

<table>
<thead>
<tr>
<th>Proposed Models</th>
<th>IRT Model Studied (Type of Item)</th>
<th>DIF Detection Method</th>
<th>The main difference from the other studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tutz and Berger (2016)</td>
<td>Rasch Model (Binary)</td>
<td>Item-focused tree based on Rasch Model but fundamentally different</td>
<td>Being able to determine the items which are responsible for DIF</td>
</tr>
<tr>
<td>Berger and Tutz (2016)</td>
<td>2PL IRT Model (Binary)</td>
<td>Item-focused tree based on Rasch Model but fundamentally different</td>
<td>Nonuniform and uniform DIF; being able to determine the items which are responsible for DIF</td>
</tr>
<tr>
<td>Jeon and De Boeck (2016)</td>
<td>2PL (Binary)</td>
<td>Logistic regression approach</td>
<td>Decomposition of item (See the article for further explanation)</td>
</tr>
<tr>
<td>Bollmann et al. (2016)</td>
<td>Partial Credit Model (Polytomous)</td>
<td>The same rationale with Tutz and Berger (2016)</td>
<td>Nonuniform and uniform DIF; being able to determine the responsible items</td>
</tr>
</tbody>
</table>

Jeon and De Boeck (2016) proposed a generalized item response tree model for psychological measurements. Unlike De Boeck and Partchev (2012) ’s study, they used two parameter IRTree model for polytomous categorical data such as Likert-type scale. They stated that the proposed model could be used to describe how item responses are distinctive with respect to a unipolar scale format. For example, the unipolar format implies only one feature of item responses such as "slightly agree", and "not agree", and bipolar scale format, which has two different elements of item responses such as "somewhat happy", and "somewhat sad". They
compared different type of models to each other regarding the dimensionality. The objective was not only to examine the characteristics of three-point and four-point Likert scale, but also to examine omission behaviors on the items. Moreover, they investigated if covariates could be utilized to predict heterogeneity in item parameters across persons. They found that putting covariates into models can be used to search for DIF. They pointed out that if node specific item parameters are different across nodes defined by a particular covariate, then the measurement is variant across nodes, indicating DIF.

Tutz and Berger (2016) examined the presence of DIF in dichotomous items by using item-focussed trees (IFT). In contrast to Strobl et al. (2015), they proposed the method using recursive partitioning techniques, known as trees, to detect the items that were responsible for DIF. The tree method was used at the item level. They found that item-focused trees helped to identify variables that were responsible for DIF as well as the interaction of variable and item on the detection of DIF. The simulation studies showed that the given fitting procedure was very useful regarding selection performance and estimation accuracy. Moreover, the differences between the presented method and method (Rasch Tree) proposed in Strobl et al. (2015) were given in the study (see Tutz and Berger (2016) for further information). Generally speaking, they stated that the Rasch Tree method does not detect items which are responsible for DIF but it still provides the advantage of the recursive partitioning algorithm (tree), which means that there is no need to prespecify the groups for numeric variables.

Berger and Tutz (2016) proposed an alternative method combining trees and logistic regression methodology to investigate nonuniform (NUDIF) and uniform (UDIF) in a nonparametric way, which was named as the IFT in this study. They used the advantages of the algorithm presented by Tutz and Berger (2016). They demonstrated that the method was useful
to be included in several covariates on different scales such as ordinal and continuous covariates. They also compared the extended logistic model using to investigate DIF to the proposed trees for both types of DIF. The IFT exhibited the same performance as the classical approach in the setting where two groups were present. In addition to that, the proposed method was found useful in more complex settings, such as including several covariates due to the flexibility and interpretability of IFT. The results were obtained by the R add-on package Diftree (Berger, 2015).

Bollmann, Berger, and Tutz (2016) `s study used an item-focussed tree based on the PCM for the identification of items that induced DIF, which was called the PCM-I FT. The identification was carried out with the proposed recursive partitioning method in Tutz and Berger (2016). The difference between the method (called TREE-PCM) proposed by El-Komboz et. al (2014) (see Table 2.1 and a summary of this study below) and the presented study was that the proposed method was able to detect of items which was responsible for DIF but TREE-PCM was able to detect the regions of the mode covariate that were associated with DIF. They found that for the first simulation study with one binary covariate, the PCM-I FT had higher false and true positive rates on the covariate level across DIF strengths as compared to the TREE-PCM. The comparison of the detection rates was not made on the item level since the TREE-PCM did not identify single item that was responsible for DIF. Also, the second simulation study with three different covariates showed very similar results like in the first simulation study. As per the simulation study with non-homogenous DIF, true and false positive rates on the item level for the PCM-I FT was higher like in the first simulation study. They pointed out that an alternative procedure could be to examine DIF in single thresholds.
Testing DIF Hypotheses at the Test Level

The method proposed by Strobl et al. (2015), called as Rasch Tree, was used in three studies.

Table 2-2 shows the features of the studies.

Table 2-2. The method testing DIF at the global level

<table>
<thead>
<tr>
<th>Proposed Models</th>
<th>IRT Model Studied (Type of Item)</th>
<th>DIF Detection Method</th>
<th>The main difference from the other studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strobl et al. (2015)</td>
<td>Rasch Model (Binary)</td>
<td>Based on the structural change approach to determine the changes in item parameters</td>
<td>At global test level, not detecting an item which is responsible for DIF</td>
</tr>
<tr>
<td>El-Komboz et al. (2014)</td>
<td>Rating Scale and Partial Credit Model (Polytomous)</td>
<td>Based on the structural change approach to determine the changes in item parameters</td>
<td>Similar with Strobl et al. (2015), only difference in response formats</td>
</tr>
</tbody>
</table>

Apinyapibal, Lawthong, and Kanjanawasee (2015) compared the performance of three DIF detection methods of dichotomously scored items. These methods are logistic regression based on classical test theory, SIBTEST based on item response theory, and Rasch Tree based on recursive partitioning algorithm. The study varied simulation conditions by sample size, number of DIF items, and number of items. They pointed out some similarities and differences among the three methods. They stated that the Rasch tree method can identify not pre-defined groups; therefore, the results can be investigated by testing the test groups. Also, they implied that the model-based recursive partitioning method can be thought of classification tree, which shows
similar power and Type 1 error rates as those obtained from the Manteal-Heanszel and logistic regression methods on the detection of DIF. They found that the Rasch tree method has parallel limitations to the logistic regression.

El-Komboz, Zeileis, and Strobl (2014) proposed a model-based recursive partitioning algorithm with IRT framework to detect DIF and differential step functioning (DSF) in polytomous items. For this purpose, the rating scale and partial credit model were used to estimate the parameters. A comparison was made between the proposed method and the well-established likelihood ratio test (LRT) for investigating DIF. The results showed that with respect to a given significance level, the proposed method had very similar results in the case in which simple DIF groups were available as compared to the LRT. Also, they indicated that the results of method showed higher power than those of the LRT in the case of complex DIF groups. Moreover, the power of the proposed method for the detection of DSF was slightly lower than the LRT under the RSM and PCM. For some DSF patterns, the proposed method based on the RSM has lower power than the proposed method based on the PCM.

Strobl, Kopf, and Zeileis (2015) proposed a new method (called as Rasch Tree) for the detection of the groups that induced DIF by using the structure of Rasch tree. They found that for numeric covariates, a new method was effective to specify a cutpoint for determining focal and reference groups as well as for the interpretability due to advantages of recursive partitioning algorithm methods. This study as well as El-Komboz et. al. (2014) demonstrated that their proposed tree-based DIF methods did not require practitioners to prespecify the groups in advance for numeric covariates, which is similar with the other tree-based DIF methods covered in this section such as IFT DIF method (e.g., Tutz and Berger, 2016). They reported that the proposed method in both studies carries a high power to detect DIF by examining item instability.
based on structural change test. They defined four steps to determine the tree structure, with Figure 2-4 as an example. Due to the positive results of these two studies, the tree-based method based on the Rasch model for DIF detection that will be studied in this paper was proposed and used in Strobl et. al. (2015) and El-Komboz et. al. (2014). This method will henceforth be referred to as the Rasch tree method for DIF detection.

![Figure 2-4](image)

Figure 2-4. An example of the Rasch tree model obtained in the present study with one binary covariate

**Research Questions**

As demonstrated above, IRTree methods have proven useful for DIF detection in item response data and have also been proven useful for handling missing item response data. However, studies that tested these psychometric applications of IRT did so in isolation. For example, DIF detection was tested when no data was missing, and the treatment of missing data was tested under an assumption of no DIF. Measurement practitioners, however, often have to
deal with data that have issues such as DIF and missingness occurring within the same data set.

The purpose of this study is to examine the performance of the Rasch tree global hypothesis DIF detection method when missingness and DIF are concurrently present in datasets. The research questions are:

1) How well does the Rasch tree DIF detection method work (as evaluated by Type I error rates and power) when large DIF is present in a single item, various percentages of missing data are present in the DIF item, and missing data are deleted listwise?

2) How well does the Rasch tree DIF detection method work (as evaluated by Type I error rates and power) when large DIF is present in a single item, various percentages of missing data are present in the DIF item, and missing data are deleted listwise?
CHAPTER 3
METHOD

Simulation Method and Settings

In order to answer the research questions, a simulation study was conducted by using the R.3.3.2 program (R Development Core Team, 2016), using add-on package psychotree (Zeileis, Strobl, Wickelmaier, & Kopf, 2012) for the tree-based IRT model used by Strobl et. al. (2015) and El-Komboz et. al. (2014). The full R codes for the simulation study are presented in Appendix.

Sample Size and Number of Items

1000 replications for each simulation conditions were used to obtain an accurate result for the estimates of Type I error and the power of the method for the detection of DIF. The total sample size was 1000 in this study, which was 500 for each focal and reference group. The selected sample size condition can be thought as an enough amount for DIF detection (e.g., Robitzsch & Rupp, 2009; Finch, 2011a, 2011b; El-Komboz, Zeileis & Strobl, 2014). 20 multiple choice items were utilized for the current study, which respondents should select either right or wrong answer (i.e., binary test data). One of the twenty items was simulated to have DIF, making it the target item for the DIF analysis. The choice to simulate DIF into only one item was consistent with some (not all) other DIF studies (e.g., Robitzsch & Rupp, 2009; Finch, 2011a, 2011b).

Person Parameters

The ability parameters were generated from a standard normal distribution $N(0, 1)$ within each of the groups.
**Item Parameters**

The Rasch model was used for all data generation. Item difficulty parameters for both reference and focal groups were generated from a standard normal distribution $N(0, 1)$ in order to obtain a range of item difficulty that would provide information at the various ability levels of the simulees. For items that did not have DIF effects, the same item parameters were used across groups to generate all item responses. For items exhibiting DIF, different item parameters were used for the data generating process. Specifically, the value 1 was subtracted from the first item’s difficulty parameter to obtain a unique item parameter for the reference group. Therefore, the item was easier for the reference group as compared to the focal group. This is considered large DIF under ETS classification of DIF magnitudes (Zwick, Thayer, & Lewis, 1999).

**Predictor Variables**

In the present study, a binary predictor variable was used to define the reference and focal group. The binary predictor was categorized into two parts: reference and focal group. It was randomly selected to be placed into datasets. The first 500 of which was selected for the reference group and the rest was picked for the focal group.

**Percentages of Missing Data**

The percentages of missing data for the target item, and the second item which does not exhibit DIF, was created at the level of 5%, 10%, 15%, 20%, 25%, and 30%. These amounts of missing responses are consistent with those used in related studies (e.g., Robitzsch & Rupp, 2009; Finch, 2011a, 2011b; Emenoglu, Falenchuk & Childs, 2010). Missingness was placed into both a DIF item and a non-DIF item to allow for different types of evaluation criteria to be computed (i.e., to provide information about Type I errors and Type II errors in the presence of missingness).
**Type of Missing Data**

For missing responses, MCAR, MAR, and MNAR mechanisms were simulated to remove data and create missingness. These types of missing mechanisms were used in prior studies (e.g., Robitzsch & Rupp, 2009; Finch, 2011a, 2011b; Emenoglu, Falenchuk & Childs, 2010). Based on the idea in Acock’s (2005) study, all missing mechanisms were applied to the current study. For MCAR, examinees were randomly selected to have their item 1 data removed. An equal number of examinees were selected from both groups.

In the case of the MAR condition, the goal was to make the missing data relate to the group membership. We achieved this by allocating 2/3 of the missing data to examinees who belonged to a specific group such as the reference group of the model covariate. Then, the remaining 1/3 of missing data were put on the target item by selecting those who belonged to the reference group of the model covariate at the same time (Finch, 2011b). That is, so as to make this condition more realistic, missing responses were placed into two groups but mostly went into the reference group rather than putting missingness into only the reference group in contrast to Finch (2011b).

As per MNAR data, two different MNAR conditions were created. In the case of the MNAR1 condition, the goal was to make the missing data relate to the item data itself (i.e., an equal proportion of examinees who scored correctly and examinees who scored incorrectly on the DIF item were sampled). We achieved this by allocating 2/3 of the missing data to examinees who correctly answered the target item across both groups. Then, the remaining 1/3 of missing data were put on the target item by selecting those who got the target item wrong across both groups (e.g., Finch, 2011b).

For MNAR 2 condition, 4/9 of missing data in the given percentages was selected from those in the reference group who incorrectly selected the target item, and put them on the target
item. 2/9 of missing responses was replaced with those in the reference group who gave the right answer to the target item. 2/9 of missing data in the given percentages was selected from those in the focal group who wrongly answered the target item, and put them on the target item. 1/9 of missing responses was replaced with those in the focal group who picked the right answer for the target item. (Rose, Von Davier & Nagengast, 2015). Subsequently, they were sampled from those selected observations. For the condition where no DIF item was simulated, the same idea was followed.

**Analyses**

Before estimating the model parameters, data preparation should be made for further analysis. Strobl, Kopf, and Zeileis (n.d.) stated that if there are the cases that do not contribute to the Rasch model since they do not give any information about the item characteristics, those cases should be excluded from the dataset. Those cases in which all rows include either 0 or 1 should be removed. Then the cases are selected for further analysis if the examinee does not answer all items either correctly or wrongly. Furthermore, due to the nature of the Rasch tree, missing data are removed from the dataset for further analysis, which means that the method treats missing data as listwise deletion method does. This missing treatment method appears a default for the psychotree package. After data are ready for the analysis, the conditional maximum likelihood approach is used to estimate item parameters. After estimating the item parameters, the individual deviations appear concerning a covariate in the model. When DIF is present between subgroups of the covariate, the variations will regularly be seen. On the other hand, the values will appear at random without the existence of DIF (Strobl, Kopf, & Zeileis, 2015; El-Komboz, Zeileis, & Strobl, 2014).

Once a covariate has been picked for splitting, the cutpoint is identified by maximizing the partitioned log-likelihood within the range of the covariate (Strobl et al., 2015). For binary
covariate, since there is only one obtainable split, the selection of cutpoint is not as important as the conditions in which the numeric covariate is present. After selecting the cutpoints, two different stopping criteria are used to implement in this step. For the first one, the tree continues splitting when significant parameter instability is identified on the covariates. The significance level usually sets to 5% to be used as the stopping criteria. The last one depends on checking the optimal sample size per node. Someone might want to see Strobl et al. (2015) for details on the method.

**Evaluation Criteria**

The study evaluates whether the method correctly detects our binary covariate causing DIF in the presence of missing item responses since the method is based on the recursive partitioning approach, which splits the covariate space in terms of the differences in item parameters. The method tests a null hypothesis about DIF at the test level rather than at the item level. Specifically, the Rasch tree method is testing the null hypothesis of $H_0: b_r = b_f$ for all items. Here, the null hypothesis can only be satisfied if all items are DIF free in $b$ parameters, and the alternative hypothesis is supported if one of any of the items displays DIF. This test-level null hypothesis is evaluated as opposed to an item-level null hypothesis because the Rasch tree DIF method requires fitting separate parametric models in each covariate spaces (see Strobl et al., 2015; El-Komboz et al, 2014; Berger & Tutz, 2016; Tutz & Berger, 2016).

**Type 2 Error Rate**

In the simulation study of obtaining Type 2 error rates, Item 1 was determined as a target item for DIF detection, which means that item parameters for Item 1 included a prespecified difference between the reference and focal group. Six different percentages of missing data were put in the target item to detect their impacts, and to compare one to another (5%, 10%, %5%, 20%, 25%, and 30%). Power could be calculated as $power = 1-Type II error rate$, however these
results are redundant with the Type II error results and hence are not directly presented in the result tables.

**Type 1 Error Rate**

To find Type 1 error rates, another simulation study was conducted for non-DIF condition. For this condition, the same amounts of missing data mentioned above were put into Item 1. The results for Item 1 including missingness but no DIF was used to obtain Type 1 error rates for various missing data mechanisms and amounts. The results for both evaluation criteria will be given below.
CHAPTER 4
RESULTS

Type 1 Error Rate

Type 1 error rates by the Rasch tree method, amount of missing responses, and types of missing mechanism appear in Table 4-1. Although the Type 1 error rates were all close to the nominal level, the slight differences across all conditions might be mentioned.

Figure 4-1 shows that there was an interaction between percentages of missing data and missing data type in their effects on Type I error rates of the Rasch tree DIF detection method. More specifically, when data were simulated to be MNAR2, 15%, and 30% missing case displayed higher Type 1 error rate than the others with the same proportions. Also, it seemed that when datasets included 15% of missingness with the MNAR1 mechanism, the method showed somewhat highest error rate across all conditions. For the condition in which the data had 10% of missing responses with MCAR mechanism, the method had slightly higher rate than those with the same amount of missingness. For 20% of missingness, the method had very similar Type 1 error rate of all types of missing mechanisms. In the case where data were simulated to be MNAR1, MAR, and MNAR2 with 15%, and 20%, the Type 1 error rates declined as percentages increased. As per the datasets which included 25% of missingness, and were generated to be MNAR1, Type 1 error rate was slightly underestimated for the Rasch tree method. Finally, when the data were simulated to be MNAR2 with 30% missing case, the method had slightly higher Type 1 error rates than the others. On the other hand, an increasingly Type 1 error rates across percentages of missing data for each type of missing data mechanisms were not found in Table 4-1 and Figure 4-1.
Type 2 Error Rate or Power

Type 2 error rates are demonstrated in Table 4-1 for the case of missing data that were created only in Item 1. The results for all conditions indicated that the method had power well above the desired 0.80, as the power of the method was often substantially greater than 0.90.

A main effect of percentages of missing data can be seen Figure 4-2, in that it was most often the case that higher proportions of missing data resulted in less power across the different missingness types. However, some interaction type effects were observed. More specifically, the results demonstrated that when data were MCAR, and MAR, the power rates declined as the proportion of missing data increased. For the MNAR2, the power rates were somewhat lower when the proportions of missing data were 20%, 25%, and 30% as compared to the other missing percentages, and the power increased as missing data percentages increased from 20% to 25%. Furthermore, when data included 15%, and 20% of missing cases, MNAR1 showed lower power rates for the method. On the other hand, the power rates showed a decreasing slope for the MCAR, and MAR when the percentages of missing data increased from 5% to 30%.
Table 4-1. Type 1 and Type 2 error rates for the Rasch tree method by percentages of missing data, and type of missing data

<table>
<thead>
<tr>
<th>Percentages of Nonresponses</th>
<th>Missing Mechanism</th>
<th>Type 1 Error</th>
<th>Type 2 Error (1-Power)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>MCAR</td>
<td>0.047</td>
<td>0.028</td>
</tr>
<tr>
<td>10%</td>
<td>MCAR</td>
<td>0.045</td>
<td>0.030</td>
</tr>
<tr>
<td>15%</td>
<td>MCAR</td>
<td>0.037</td>
<td>0.038</td>
</tr>
<tr>
<td>20%</td>
<td>MCAR</td>
<td>0.037</td>
<td>0.066</td>
</tr>
<tr>
<td>25%</td>
<td>MCAR</td>
<td>0.048</td>
<td>0.074</td>
</tr>
<tr>
<td>30%</td>
<td>MCAR</td>
<td>0.048</td>
<td>0.087</td>
</tr>
<tr>
<td>5%</td>
<td>MAR</td>
<td>0.052</td>
<td>0.026</td>
</tr>
<tr>
<td>10%</td>
<td>MAR</td>
<td>0.047</td>
<td>0.044</td>
</tr>
<tr>
<td>15%</td>
<td>MAR</td>
<td>0.051</td>
<td>0.046</td>
</tr>
<tr>
<td>20%</td>
<td>MAR</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>25%</td>
<td>MAR</td>
<td>0.043</td>
<td>0.082</td>
</tr>
<tr>
<td>30%</td>
<td>MAR</td>
<td>0.042</td>
<td>0.103</td>
</tr>
<tr>
<td>5%</td>
<td>MNAR1</td>
<td>0.065</td>
<td>0.034</td>
</tr>
<tr>
<td>10%</td>
<td>MNAR1</td>
<td>0.048</td>
<td>0.036</td>
</tr>
<tr>
<td>15%</td>
<td>MNAR1</td>
<td>0.044</td>
<td>0.031</td>
</tr>
<tr>
<td>20%</td>
<td>MNAR1</td>
<td>0.052</td>
<td>0.032</td>
</tr>
<tr>
<td>25%</td>
<td>MNAR1</td>
<td>0.038</td>
<td>0.06</td>
</tr>
<tr>
<td>30%</td>
<td>MNAR1</td>
<td>0.052</td>
<td>0.067</td>
</tr>
<tr>
<td>5%</td>
<td>MNAR2</td>
<td>0.049</td>
<td>0.028</td>
</tr>
<tr>
<td>10%</td>
<td>MNAR2</td>
<td>0.048</td>
<td>0.030</td>
</tr>
<tr>
<td>15%</td>
<td>MNAR2</td>
<td>0.062</td>
<td>0.036</td>
</tr>
<tr>
<td>20%</td>
<td>MNAR2</td>
<td>0.047</td>
<td>0.055</td>
</tr>
<tr>
<td>25%</td>
<td>MNAR2</td>
<td>0.053</td>
<td>0.053</td>
</tr>
<tr>
<td>30%</td>
<td>MNAR2</td>
<td>0.059</td>
<td>0.065</td>
</tr>
</tbody>
</table>
Figure 4-1. Type 1 error rates across all conditions
Figure 4-2. Power rates across all conditions
CHAPTER 5
DISCUSSION

The results presented in the present study provide the answers to our research questions. In the first research question, we asked whether the proportion of missing data would have an impact on the performance of the Rasch tree method for the detection of DIF. The power of the method for the detection of DIF declined as the proportion of missing data increased, but all the power results were indicative of a DIF detection method that has acceptable to strong statistical power. Also, the decreases apparently appeared when data were simulated to be MCAR, and MAR. Finch (2011a)’s study indicated that power for LR method with listwise deletion method had a decreasing tendency as the missingness increased, which was very similar with our results when the Rasch tree method was used.

For the Type 1 error rates, the results showed that with %5 missing case, the error rate for the MNAR 1 was highest as compared to the results for higher proportions; however, when data were simulated to be MAR with 30% of missing data, the Type 1 error rate was lowest as compared to the others. Similarly, Finch (2011a) found that when data were MAR2 (those who had lower total exam scores randomly selected for missing case), LR method with listwise showed the lowest Type 1 error rate in the highest proportion of missing data. In addition to that, listwise deletion treatment for missing data had an inflated error rate for logistic regression across all levels of missing data in the case of MAR1 (those in the focal group randomly selected for missingness). Also, when data were MNAR, in contrast to our results, Finch (2011a)’s study indicated that all DIF detection methods with listwise deletion method had a decreasing error rate as the missing data increased. It was also important to note that in Finch (2011a)’s study, the Type 1 error rate for three DIF methods was the lowest when data were MNAR (only those who did not score correctly was taken) with %30 missing case. In our study, the similar results were
seen in the case of both MNAR conditions with %30 missing case. On the other hand, Finch (2011b) found no differences in the power rates of three methods when the proportion of missingness increased.

Our final research question asked to what degree the Rasch tree method detected DIF when different types of missing data varied in the data set. The results showed that when data were MCAR, and MAR, the error rates for the method were very similar and near the nominal level of 0.05; however, MNAR1 mechanism with %5 inflated Type 1 error rate. Finch (2011a, 2001b) showed that Type 1 error rates were near the nominal level across all types of missing data. Similarly, Finch (2011a) found that when data were simulated to be MNAR (the same rationale with our MNAR1), the smallest proportion (10%) showed the highest Type 1 error rate when three methods with listwise deletion were used. Also, when the percentages of missing data increased from 10% to 30%, the error rates of three methods with listwise deletion showed a decreasing slope, which was somewhat different from our results. In Figure 4-1, no systematic changes were explicitly appeared with respect to the mechanisms. Additionally, Finch (2011a)`s study demonstrated that there were no obvious differences in the Type 1 error rates across all mechanism with listwise deletion method for missing data, though it was very close to the nominal level.

The power of the method for the detection of DIF regarding the second research question was high across all simulated conditions. However, the power rates for the method were lowest when data were MCAR with 20%, 25%, and 30% of missing cases. Also, when all missing mechanisms was the case, the power declined as the proportion of missing data increased but both MNAR mechanism had higher power rates than that of both MAR, and MCAR. Finch (2011a) found that power rates decreased as the amount of missingness increased, and the result
also showed that power rates for all methods increased as the overall sample size increased. Finch (2011b)’s result showed that when data were MCAR, the power rates of three different methods (LR, M-H, and SIBSET) with listwise deletion were slightly lower than other mechanisms (MAR, and MNAR).

In this study, there was only one DIF magnitude, which was considered as a large DIF according to the ETS system for DIF classification (Zwick, 2012). In general, all DIF methods would detect large, uniform DIF with relatively high power and low Type I and Type II error rates, as compared to the detection of small or moderate DIF. For this reason, several magnitudes such as small, moderate, and large DIF should be used to compare and to test the performance of the Rasch tree method on the detection DIF.

Moreover, we used conditions in which only missing responses were placed in the target item. The reason for doing this was that the Rasch tree method treated missing data as listwise deletion, which might lead to a significant amount of loss in the observations in the data sets. We compared the results in the present study with those obtained in the related studies using different DIF detection methods such as LR, and using various missing methods such as listwise deletion method for missing data. Also, our results are not comparable with those results which the model-based recursive partitioning algorithm was used under the IRT structure, but those models are based on different DIF detection methods since they did not focus on the DIF detection in the presence of missing data.

Another limitation of the study is that we used only one covariate to identify DIF, which was binary variable. That is, the study did not utilize the advantage of the method which indicates that there is no more need to prespecify the cutpoints for the numeric covariates. In contrast, the future study consisting of both numeric and categorical might be used for the
comparison with the current study which included several missing mechanisms. Next limitation is about the sample size condition. We used an even sample size of the subgroups; however, the datasets collecting for the analysis in the real settings may contain unequal sample sizes of the subgroups. That is, the sample size and, inequality between the sample size of subgroups might have an impact on the performance of the DIF detection method since the splitting criteria require well enough sample size for each subgroup. In addition to that, as compared to the others, the number of iteration might be increased from 1000 to at least 2500 with all additional conditions.

The next limitation would be thought that the study evaluated DIF between two groups with the same latent ability. It is important that future research examines the Rasch tree DIF method in the presence of both missing data and group impact (i.e., differences in latent ability between groups). Another limitation is that the listwise deletion was the default for the Rasch tree method, removing the cases from the datasets. However, it might be possible to use different types of methods for handling the missing data, and the performance of the methods might be investigated across various kinds of missing data treatments.

Finally, it is important to note that in the data generating process, the amount of missing data with the same percentages across all mechanisms did not necessarily result in the same amount of missing data when viewing the item data in total. Therefore, the comparisons that were made here might be affected by any confounding factors of total percentages of missing data in the item. However, since MAR and MCAR were based on missingness definitions that result in the same amount (i.e., \( n \)) of missing cases, the comparisons between both conditions were probably more plausible.
APPENDIX A
SIMULATION CODES FOR MCAR MECHANISM

```r
require(pscyhotree)
flagmcar = c()
totmcar= c()

# start missing percentages loop around here
for(perc in c(.05, .1, .15, .2, .25, .3)){

# start 1000 iteration loop around here
for (iter in c(1:1000)){

##############################################################
############      Generation            ######################
##############################################################

g1_ss <- 500
g2_ss <- 500
ss <- g1_ss+g2_ss
numitems <- 20

# creating two groups of simulees
g1_theta <- rnorm(g1_ss,0,1) # change sample size or distribution parameters as desired
g2_theta <- rnorm(g2_ss,0,1)

# create a test of 20 items
b <- rnorm(numitems,0,1)
iparam <- data.frame("item"=c(1:numitems),"b"=b)

g1_iparam <- iparam
g2_iparam <- iparam
g2_iparam[1,2] <- g2_iparam[1,2]-1 # creating uniform DIF in item 1 with magnitude of -1
(bg2-bg1), to be varied in simulation
g2_iparam[1,2] <- g2_iparam[1,2]    # No DIF
# specifying rasch model

rm <- function(iparam, theta, i){
  ((exp((theta-iparam[i,2])))/
   (1+(exp(theta-iparam[i,2]))))
}

# creating empty matrix to hold item responses
g1_data <- c()
g2_data <- c()

# Generating item responses g1
for(j in 1:numitems) {
  g1_data <- cbind(g1_data,
```
ifelse(rm(g1_iparam,g1_theta,j) >= runif(length(g1_theta)),1,0))
}

# Generating item responses g2
for(j in 1:numitems) {
  g2_data <- cbind(g2_data,
  ifelse(rm(g2_iparam,g2_theta,j) >= runif(length(g2_theta)),1,0))
}

data_all <- rbind(g1_data,g2_data)
group <- c(rep(1,g1_ss),rep(2,g2_ss))
id <- c(1:ss)
data_all <- cbind(data_all,id,group,cov)
Y3mcar <- as.data.frame(data_all)

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ data generation done ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Enter Missing ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# putting missingness into data
data3= subset(Y3mcar, select= c(V1,id))
for(j in 1:1){ # opening item loop to create missing item by item
  r0_missymcar <- sample(data3$id, perc*ss) # missing is MCAR (make it something else)
  # putting missing in item 1 for people who originally scored 0 on item 1
  for(p in 1:ss){ # person loop to create missing in one item, person by person
    for(r in r0_missymcar){ # loop to compare each rand value to each person's id
      if(Y3mcar[p,(numitems+1)]==r){
        Y3mcar[p,j]<- 99
      }
    } # closing missy loop
  } # closing person loop
}
data33= subset(Y3mcar2, select= c(V2,id))

#### Data Preparation ####
Y3mcar$group <- as.character(Y3mcar$group)
Y3mcar$group[Y3mcar$group==1] = "male"
Y3mcar$group[Y3mcar$group==2] = "female"
Y3mcar$group = as.factor(Y3mcar$group)

mydata_mcar=Y3mcar
mydata_mcar$resp <- as.matrix(mydata_mcar[,1:20])
mydata_mcar <- mydata_mcar[,-(1:20)]
mydata_mcar=mydata_mcar[-1]
mydata_mcar$resp[mydata_mcar$resp==99]=NA
mydata_mcar <- subset(mydata_mcar, rowMeans(resp, na.rm = TRUE) > 0 & + rowMeans(resp, na.rm = TRUE) < 1)

######################################################################### end of entering missingness

######################################################################### Analysis
#########################################################################
model1mcar <- raschtree(formula= resp ~ group, data = mydata_mcar)
itemmcar=as.data.frame(itempar(model1mcar))
coefmcar=as.data.frame(coefficients(model1mcar))

if (length(model1mcar) > 1) {flag1=1} else {flag1=0}

flagmcar= c(flagmcar, flag1)
vector <- c(iter,perc,flag1)
totmcar <- rbind(totmcar,vector)

} #iteration loop
} # percentages of missingness loop
# putting missingness into data
data3 = subset(Y3mar1, select= c(V1, id, group))
c3mar1 = data3$id[data3$group==1]
f3mar1 = data3$id[data3$group==2]

for(j in 1:1){ # opening item loop to create missing item by item
    perc.3 <- perc/3  # dividing the total perc of missing into thirds
    r0_perc <- perc.3*2  # giving 2/3 person of missingness
    r1_perc <- perc.3  # giving 1/3 person of missingness

    r0_missy <- sample(c3mar1, r0_perc*1000)  # missing is MAR
    r1_missy <- sample(f3mar1, r1_perc*1000)

    # putting missing in item 1 for people who originally scored 0 on item 1
    for(p in 1:ss){ # person loop to create missing in one item, person by person
        for(r in r0_missy){ # loop to compare each rand value to each person's id
            if(Y3mar1[p,(numitems+1)]==r){
                Y3mar1[p,j] <- 99
            }
        }
    }

    # putting missing in item 1 for people who originally scored 1 on item 1
    for(p in 1:ss){ # person loop to create missing in one item, person by person
        for(r in r1_missy){ # loop to compare each rand value to each person's id
            if(Y3mar1[p,(numitems+1)]==r){
                Y3mar1[p,j] <- 99
            }
        }
    }

    # Data Preparation
    Y3mar1$group <- as.character(Y3mar1$group)
Y3mar1$group[Y3mar1$group==1] = "male"
Y3mar1$group[Y3mar1$group==2] = "female"
Y3mar1$group = as.factor(Y3mar1$group)

mydata_mar1=Y3mar1
mydata_mar1$resp <- as.matrix(mydata_mar1[, 1:20])
mydata_mar1 <- mydata_mar1[, -(1:20)]
mydata_mar1=mydata_mar1[-1]
mydata_mar1$resp[mydata_mar1$resp==99]=NA
mydata_mar1 <- subset(mydata_mar1, rowMeans(resp, na.rm = TRUE) > 0 &
  + rowMeans(resp, na.rm = TRUE) < 1)

############################# end of entering missingness ###########################

#################################################################
############     Analysis                ######################
#################################################################
model1mar1 <- raschtree(formula= resp ~ group,
  data = mydata_mar1)

totmar1 <- rbind(totmar1,vector)
}
} # percentages of missingness loop

itemmar1=as.data.frame(itempar(model1mar1))
coefmar1=as.data.frame(coefficients(model1mar1))

if (length(model1mar1) > 1) {flag1=1} else {flag1=0}
flagmar1= c(flagmar1, flag1)

vector <- c(iter,perc,flag1)
totmar1 <- rbind(totmar1,vector)
} #iteration loop

} # # percentages of missingness loop
APPENDIX C
SIMULATION CODES FOR MNAR MECHANISM

# putting missingness into data
data3= subset(Y3mnar1, select= c(V1,id))
c3mnar1=data3$id[data3$V1==0]
f3mnar1=data3$id[data3$V1==1]
c3m1=length(c3mnar1)
f3m1=length(f3mnar1)

for(j in 1:1){ # opening item loop to create missing item by item
    perc.3 <- perc/3  # dividing thh total perc of missing into thirds
    r0_perc <- perc.3*2 #giving 2/3 person of missingness
    r1_perc <- perc.3 #giving 1/3 person of missingness

    r0_missymnar1 <- sample(c3mnar1, perc*c3m1) # missing is MNAR1
    r1_missymnar1 <- sample(f3mnar1, r1_perc*f3m1)

    #putting missing in item 1 for people who originally scored 0 on item 1
    for(p in 1:ss){ # person loop to create missing in one item, person by person
        for(r in r0_missymnar1){ # loop to compare each rand value to each person's id
            if(Y3mnar1[p,(numitems+1)]==r){
                Y3mnar1[p,j]<- 99
            }
        }
    }

    #putting missing in item 1 for people who originally scored 1 on item 1
    for(p in 1:ss){ # person loop to create missing in one item, person by person
        for(r in r1_missymnar1){ # loop to compare each rand value to each person's id
            if(Y3mnar1[p,(numitems+1)]==r){
                Y3mnar1[p,j]<- 99
            }
        }
    }
}

for(j in 1:1){ # opening item loop to create missing item by item
    perc.3 <- perc/3  # dividing thh total perc of missing into thirds
    r0_perc <- perc.3*2 #giving 2/3 person of missingness
    r1_perc <- perc.3 #giving 1/3 person of missingness

    r0_missymnar1 <- sample(c3mnar1, perc*c3m1) # missing is MNAR1
    r1_missymnar1 <- sample(f3mnar1, r1_perc*f3m1)

    #putting missing in item 1 for people who originally scored 0 on item 1
    for(p in 1:ss){ # person loop to create missing in one item, person by person
        for(r in r0_missymnar1){ # loop to compare each rand value to each person's id
            if(Y3mnar1[p,(numitems+1)]==r){
                Y3mnar1[p,j]<- 99
            }
        }
    }

    #putting missing in item 1 for people who originally scored 1 on item 1
    for(p in 1:ss){ # person loop to create missing in one item, person by person
        for(r in r1_missymnar1){ # loop to compare each rand value to each person's id
            if(Y3mnar1[p,(numitems+1)]==r){
                Y3mnar1[p,j]<- 99
            }
        }
    }
}
### Data Preparation ###

```r
Y3mnar1$group <- as.character(Y3mnar1$group)
Y3mnar1$group[Y3mnar1$group==1] = "male"
Y3mnar1$group[Y3mnar1$group==2] = "female"
Y3mnar1$group = as.factor(Y3mnar1$group)
mydata1=Y3mnar1
mydata1$resp <- as.matrix(mydata1[, 1:20])
mydata1 <- mydata1[, -(1:20)]
mydata1=mydata1[, -1]
mydata1$resp[mydata1$resp==99]=NA
mydata1 <- subset(mydata1, rowMeans(resp, na.rm = TRUE) > 0 &
    + rowMeans(resp, na.rm = TRUE) < 1)
```


```
###########################################################################
###########################################################################
###########################################################################
Analysis  ###########################################################################
###########################################################################
modelmnar1 <- raschtree(formula= resp ~ group,
    data = mydata1)
#extracting item parameters for each node
itemmnar1=as.data.frame(itempar(modelmnar1, node=2))
#extracting coefficients for each node
coefmnar1=as.data.frame(coefficients(modelmnar1, node=2))

if (length(modelmnar1) > 1) {flag1=1} else {flag1=0}
flagmnar1= c(flagmnar1, flag1)
vector <- c(iter, perc, flag1)
totmnar1 <- rbind(totmnar1, vector)

} #iteration loop
} # percentages of missingness loop
```
APPENDIX D
SIMULATION CODES FOR MNAR2 MECHANISM

# putting missingness into data
data3 = subset(Y3mnar2, select= c(V1, id, group))

c3 = data3$id[data3$V1 == 0 & data3$group == 1]
f3 = data3$id[data3$V1 == 1 & data3$group == 1]
f33 = data3$id[data3$V1 == 1 & data3$group == 2]
c33 = data3$id[data3$V1 == 0 & data3$group == 2]
c3m2 = length(c3)
f3m2 = length(f3)
c3m22 = length(c33)
f3m22 = length(f33)

for(j in 1:1) { # opening item loop to create missing item by item
  perc.3 <- perc / 3 # dividing the total perc of missing into thirds
  r0_perc <- perc.3 * 2 # giving 2/3 person of missingness
  r1_perc <- perc.3 # giving 1/3 person of missingness

  r0_missy <- sample(c3, (r0_perc * c3m2) * (2 / 3)) # missing is Mnar2 (make it something else)
  r0_missy1 = sample(c33, (r0_perc * c3m22) / 3)
  r1_missy1 = sample(f33, (r1_perc * f3m22) / 3)
  r1_missy <- sample(f3, (r1_perc * f3m2) * (2 / 3))

  # putting missing in item 1 for people who originally scored 0 on item 1
  for(p in 1:ss) { # person loop to create missing in one item, person by person
    for(r in r0_missy) { # loop to compare each rand value to each person's id
      if(Y3mnar2[p, (numitems + 1)] == r) {
        Y3mnar2[p, j] <- 99
      }
    }
  }

  # putting missing in item 1 for people who originally scored 1 on item 1
  for(p in 1:ss) { # person loop to create missing in one item, person by person
    for(r in r1_missy) { # loop to compare each rand value to each person's id
      if(Y3mnar2[p, (numitems + 1)] == r) {
        Y3mnar2[p, j] <- 99
      }
    }
  }
}
Y3mnar2[p,j]<- 99}

} # closing missy loop
} # closing person loop for missy

#putting missing in item 1 for people who originally scored 1 on item 1
for(p in 1:ss){ # person loop to create missing in one item, person by person

for(r in r1_missy){ # loop to compare each rand value to each person's id

if(Y3mnar2[p,(numitems+1)]==r){
   Y3mnar2[p,j]<- 99

} # closing missy loop
} # closing person loop for missy

}

for(p in 1:ss){ # person loop to create missing in one item, person by person

for(r in r1_missy1){ # loop to compare each rand value to each person's id

if(Y3mnar2[p,(numitems+1)]==r){
   Y3mnar2[p,j]<- 99

} # closing missy loop
}
}

} # closing item loop for missy

### Data Preparation ###
Y3mnar2$group <- as.character(Y3mnar2$group)
Y3mnar2$group[Y3mnar2$group==1] = "male"
Y3mnar2$group[Y3mnar2$group==2] = "female"
Y3mnar2$group = as.factor(Y3mnar2$group)
mydatamnar2=Y3mnar2
mydatamnar2$resp <- as.matrix(mydatamnar2[, 1:20])
mydatamnar2 <- mydatamnar2[, -(1:20)]
mydatamnar2=mydatamnar2[,-1]
mydatamnar2$resp[mydatamnar2$resp==99]=NA
mydatamnar2 <- subset(mydatamnar2, rowMeans(resp, na.rm = TRUE) > 0 & + rowMeans(resp, na.rm = TRUE) < 1)

########################################################################## end of entering missingness ##########################################################################

########################################################################## Analysis ##########################################################################

68
modelmnar2 <- raschtree(formula= resp ~ group,
                        data = mydatamnar2)
#extracting item parameters for each node
itemmnar2=as.data.frame(itempar(modelmnar2, node=2))

#extracting coefficients for each node
coefmnar2=as.data.frame(coefficients(modelmnar2, node=2))

if (length(modelmnar2) > 1) {flag1=1} else {flag1=0}
flagmnar2= c(flagmnar2, flag1)

vector <- c(iter,perc,flag1)
tot <- rbind(tot,vector)
} #iteration loop
} # percentages of missingness loop
LIST OF REFERENCES


Ahmet Guven was born in Izmir, Turkey. He obtained his B.A. in Science teaching/science education from Ahi Evran University at 2008. He worked as a research assistant in the Department of Science Education at the Sakarya University for three months. He gave a lecture related to the lab of the chemistry class during that time. He has won a scholarship among well-qualified people to move on his career abroad in 2013, which is supported by Turkish Government. He enrolled for a master’s program in the School of Human Development and Organizational Studies at College of Education at the University of Florida. He will receive his M.A.E in research and evaluation methodology program in August 2017.