RISE and shine: A comparison of item fit statistics for the Rasch model

Glenn Waterbury
RISE and Shine:
A Comparison of Item Fit Statistics for the Rasch Model

Glenn Thomas Waterbury

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Department of Graduate Psychology

May 2020

FACULTY COMMITTEE:

Committee Chair:  Dr. Christine DeMars

Committee Members:

Dr. John Hathcoat

Dr. Dena Pastor
Acknowledgements

I would first like to thank my brilliant advisor, Christine DeMars. I certainly would not have been able to write this dissertation, or obtain my Ph.D., with your guidance. I have learned so much from you, in terms of statistics and measurement knowledge, the fundamentals of scientific writing, conducting research, and just how to be a graduate student. I am so appreciative of our four years together and will remember my time as your advisee quite fondly.

I would also like to thank my two other committee members, Dena Pastor and John Hathcoat. Dena, you have been a wonderful teacher to me. It was a true pleasure to get to take three courses with you. I also appreciate how much thought and care you put into reading and commenting on my dissertation. I believe my dissertation is better, and I am better too, from having learned from you the past four years. John, thank you to you as well for our time together. I appreciate the big-picture, philosophical approach you brought to my course with you and to my dissertation. You make me think in a way that is unique from the other faculty members in our program. I also really enjoyed all the random conversations we had throughout the years, you are wonderful to converse with.

Finally, I would like to thank Liz. I am so thankful that I met you during my time in the program. You were invaluable in helping me learn the ins and outs of the program, and helping me with things as simple as formatting a table in excel. I also appreciate your support of me during my time throughout the program, and specifically during my dissertation. Thanks for babysitting my simulations over the summer, and thanks for often coming into the office with me to keep me company while I was writing. I am glad we get to move to California together!
# Table of Contents

Acknowledgements.......................................................................................................................... ii  
List of Tables..................................................................................................................................... v  
List of Figures..................................................................................................................................... vi  
Abstract........................................................................................................................................... viii  

I. Chapter One: Introduction.................................................................................................................. 1  
II. Chapter Two: Literature Review....................................................................................................... 5  
   Item Response Theory.......................................................................................................................... 5  
   The Rasch Model................................................................................................................................. 7  
   Beneficial properties of the Rasch Model............................................................................................ 8  
   Importance of model-data fit with the Rasch Model........................................................................... 9  
   Residuals in IRT/Rasch Modeling........................................................................................................ 10  
   IRT Fit Statistics.................................................................................................................................. 12  
   Description of Infit and Outfit............................................................................................................. 15  
   Distributional Properties of Infit and Outfit....................................................................................... 17  
   Limitations of Infit and Outfit............................................................................................................. 24  
   Smoothing.......................................................................................................................................... 25  
   The Root Integrated Squared Error...................................................................................................... 29  
   Studies Investigating RISE................................................................................................................ 32  
   Research Questions............................................................................................................................. 36  

III. Chapter Three: Methodology.......................................................................................................... 39  
   Overview........................................................................................................................................... 39  
   Design............................................................................................................................................... 39  
   Simulating Item Responses.............................................................................................................. 40  
   Calculating RISE............................................................................................................................... 44  
   Calculating RISE: Smoothing the simulated data............................................................................ 45  
   Calculating RISE: Smoothing amount.............................................................................................. 47  
   Calculating RISE: Bootstrapping........................................................................................................ 48  
   Calculating Power and Type 1 Errors: Infit/Outfit and RISE............................................................ 49  
   Justification for the Simulation Design............................................................................................. 50  

IV. Chapter Four: Results...................................................................................................................... 53  
   Research Question 1.......................................................................................................................... 53  
   Research Question 1: Type 1 errors................................................................................................... 53  
   Research Question 1: Power.............................................................................................................. 55  
   Research Question 2.......................................................................................................................... 56  

  
  
  
  
  
  iii
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Question 2: Hanning</td>
<td>56</td>
</tr>
<tr>
<td>Research Question 2: Kernel</td>
<td>59</td>
</tr>
<tr>
<td>Research Questions 3 and 4</td>
<td>61</td>
</tr>
<tr>
<td>V. Chapter Five: Discussion</td>
<td>66</td>
</tr>
<tr>
<td>Research Question 1</td>
<td>67</td>
</tr>
<tr>
<td>Research Question 2</td>
<td>68</td>
</tr>
<tr>
<td>Research Question 3</td>
<td>74</td>
</tr>
<tr>
<td>Research Question 4</td>
<td>75</td>
</tr>
<tr>
<td>Limitations</td>
<td>77</td>
</tr>
<tr>
<td>Recommendations</td>
<td>78</td>
</tr>
<tr>
<td>Future Research</td>
<td>80</td>
</tr>
<tr>
<td>Tables</td>
<td>82</td>
</tr>
<tr>
<td>Figures</td>
<td>91</td>
</tr>
<tr>
<td>References</td>
<td>126</td>
</tr>
</tbody>
</table>
List of Tables

Table 1. Type 1 Error Rates for Hanning RISE........................................... 82
Table 2. Type 1 Error Rates for Kernel RISE............................................. 82
Table 3. Type 1 Error Rates for Infit and Outfit...................................... 83
Table 4. Power of Infit............................................................................. 83
Table 5. Power of Outfit........................................................................... 84
Table 6. Power of Hanning RISE on Item 44: Wavy................................. 84
Table 7. Power of Hanning RISE on Item 45: Big Dip............................... 85
Table 8. Power of Hanning RISE on Item 46: 3PL................................... 85
Table 9. Power of Hanning RISE on Item 47: 2PL................................... 86
Table 10. Power of Hanning RISE on Item 48: 4PL................................ 86
Table 11. Power of Hanning RISE on Item 49: Hyperbolic Cosine Model... 87
Table 12. Power of Hanning RISE on Item 50: Flat Middle...................... 87
Table 13. Power of Kernel RISE on Item 44: Wavy................................ 88
Table 14. Power of Kernel RISE on Item 45: Big Dip............................... 88
Table 15. Power of Kernel RISE on Item 46: 3PL................................... 89
Table 16. Power of Kernel RISE on Item 47: 2PL................................... 89
Table 17. Power of Kernel RISE on Item 48: 4PL................................... 90
Table 18. Power of Kernel RISE on Item 49: Hyperbolic Cosine Model...... 90
Table 19. Power of Kernel RISE on Item 50: Flat Middle....................... 91
Table 20. Comparing Peak Performance of Hanning and Kernel RISE to Infit and Outfit................................................................. 91
List of Figures

Figure 1. Comparison of the Rasch implied IRF and the generating IRF for item 44 ................................................................................................................................. 91
Figure 2. Comparison of the Rasch Implied IRF and the generating IRF for item 45 ................................................................................................................................. 92
Figure 3. Comparison of the Rasch Implied IRF and the generating IRF for item 46 ................................................................................................................................. 93
Figure 4. Comparison of the Rasch Implied IRF and the generating IRF for item 47 ................................................................................................................................. 94
Figure 5. Comparison of the Rasch Implied IRF and the generating IRF for item 48 ................................................................................................................................. 95
Figure 6. Comparison of the Rasch Implied IRF and the generating IRF for item 49 ................................................................................................................................. 96
Figure 7. Comparison of the Rasch Implied IRF and the generating IRF for item 50 ................................................................................................................................. 97
Figure 8. Power results for Hanning RISE on Item 44, the wavy item ................. 98
Figure 9. Power results for Hanning RISE on Item 45, the big dip item ............... 99
Figure 10. Power Results for Hanning RISE on Item 46, the 3PL item ............... 100
Figure 11. Power results for Hanning RISE on Item 47, the 2PL item ............... 101
Figure 12. Power results for Hanning RISE on Item 48, the 4PL item ............... 102
Figure 13. Power results for Hanning RISE on Item 49, the hyperbolic cosine item ................................................................................................................................. 103
Figure 14. Power results for Hanning RISE on Item 50, the flat middle item .......... 104
Figure 15. Power results for Kernel RISE on Item 44, the wavy item ............... 105
Figure 16. Power results for Kernel RISE on Item 45, the big dip item ............... 106
Figure 17. Power results for Kernel RISE on item 46, the 3PL item ............... 107
Figure 18. Power results for Kernel RISE on item 47, the 2PL item ............... 108
Figure 19. Power results for Kernel RISE on item 48, the 4PL item ............... 109
Figure 20. Power result for Kernel RISE on item 49, the hyperbolic cosine item .. 110
Figure 21. Power results for Kernel RISE on item 50, the flat middle item ........... 111
Figure 22. Comparing the power of Infit, Outfit, Peak Hanning, and Peak Kernel RISE for item 44, the wavy item ................................................................. 112
Figure 23. Comparing the power of Infit, Outfit, Peak Hanning, and Peak Kernel RISE for item 45, the big dip item ................................................................. 113
Figure 24. Comparing the power of Infit, Outfit, Peak Hanning, and Peak Kernel RISE for item 46, the 3PL item ................................................................. 114
Figure 25. Comparing the power of Infit, Outfit, Peak Hanning, and Peak Kernel RISE for item 47, the 2PL item ................................................................. 115
Figure 26. Comparing the power of Infit, Outfit, Peak Hanning, and Peak Kernel RISE for item 48, the 4PL item……………………………………………………… 116
Figure 27. Comparing the power of Infit, Outfit, Peak Hanning, and Peak Kernel RISE for item 49, the hyperbolic cosine item……………………………………… 117
Figure 28. Comparing the power of Infit, Outfit, Peak Hanning, and Peak Kernel RISE for item 50, the flat middle item………………………………………………. 118
Figure 29. Illustration of the effects of different bandwidth parameters on the item response function for item 44, the wavy item……………………………………… 119
Figure 30. Illustration of the effects of different bandwidth parameters on the item response function for item 45, the big dip item…………………………………… 120
Figure 31. Illustration of the effects of different bandwidth parameters on the item response function for item 46, the 3PL item……………………………………… 121
Figure 32. Illustration of the effects of different bandwidth parameters on the item response function for item 47, the 2PL item……………………………………… 122
Figure 33. Illustration of the effects of different bandwidth parameters on the item response function for item 48, the 4PL item……………………………………… 123
Figure 34. Illustration of the effects of different bandwidth parameters on the item response function for item 49, the hyperbolic cosine item………………. 124
Figure 35. Illustration of the effects of different bandwidth parameters on the item response function for item 50, the flat middle item…………………………… 125
Abstract

The Rasch model implies that the relation between examinee ability and the probability of correctly answering an item can be defined solely by a small set of parameters. In the case of Rasch modeling, there are only two parameters: the ability of an examinee and the difficulty of an item. When the data meet the requirements of the Rasch model, it possesses several appealing properties that distinguish it from Classical Test Theory and more complex Item Response Theory models.

However, the desirable properties of the Rasch model only exist when the data meet its strict requirements. Therefore, it is vital to check the fit of the data to the model, both the fit of the items and the examinees. The two primary fit statistics for Rasch models are Infit and Outfit. While useful statistics, they possess some inherent deficiencies. Therefore, it may be useful to supplement them with another fit statistic. One such fit statistic, which is computed and interpreted differently than Infit and Outfit, is the root integrated squared error (RISE). The purpose of this dissertation was to compare the performance of RISE, in terms of type 1 error rates and power, to Infit and Outfit. Additionally, RISE requires statistical smoothing in its computation. Therefore, an additional purpose of this dissertation was to examine the impact of smoothing amount and smoothing type on the performance of RISE.

A simulation study was conducted to examine, RISE, Infit, and Outfit. Responses to a 50 item test were generated, with 43 items that fit the Rasch model and 7 items that did not. Sample size was manipulated and had three levels: 200, 500, or 1,000 examinees. Two smoothing techniques were used: Hanning or Kernel smoothing with a Gaussian function. Within each smoothing technique, nine smoothing amounts were used.
The results showed that RISE performed similarly across smoothing techniques.

Within each smoothing technique, smoothing amount often had a drastic impact on RISE, with the best results generally associated with a low to medium amount of smoothing.

Across most of the misfitting items, Outfit and/or Infit outperformed RISE.
CHAPTER 1

INTRODUCTION

High-stakes, large-scale testing is a major component of the educational process of many countries, including the United States. A crucial aspect of any testing process is the scoring of the tests. Large-scale tests are often scored using Rasch modeling or item response theory (IRT). Use of a Rasch or an IRT model implies that an encounter between an examinee and an item can be represented by a parametric function. Specifically, the probability of a correct response to an item from an examinee is a function of an examinee’s latent ability level, often symbolized using $\theta$, and one or more characteristics of the item. In the case of the Rasch model, the only item characteristic that is modeled is the item’s difficulty; IRT models add more parameters. IRT/Rasch models can be depicted graphically using an item response function (IRF), which indicates a monotonic, non-linear relation between an examinee’s ability and the probability of correct response to an item. The exact location and shape of the IRF is determined by the characteristics of the item.

IRT/Rasch models possess several desirable properties. One of these properties is known as parameter invariance. When using an IRT model, person and item parameters are invariant, up to a linear transformation. This means that examinees’ $\theta$’s obtained from one set of items will equal their $\theta$’s obtained from a different set of items, up to a linear transformation. The same is true of item parameters. Item parameters calibrated using one sample of examinees will equal item parameters using a different sample of examinees. Thus, person and item parameters are not sample dependent, which is a major advantage of IRT over the classical test theory approach to analyzing item response data. Rasch
models have an additional desirable property: if all examinees respond to the same items, examinees with the same number-correct score will have the same estimated $\theta$.

The desirable properties of IRT models only hold when the item response data conform to the functional form of the specified IRT model. When the data do not fit the model, item and person parameters may be biased, and parameter invariance will not hold. Thus, when using any type of IRT model, it is crucial to examine the fit of the data to the model. This is especially true for the Rasch model. The Rasch model specifies the strictest functional form of any IRT model: that an encounter between a person and an item is solely a function of the person’s ability and the item’s difficulty. This stands in contrast to more complex IRT models, which can allow items to have varying levels of discrimination and which can also model the presence of correct guessing in the data. Given the Rasch model’s strict requirements, there are many ways in which item response data can misfit the model.

For both IRT and Rasch modeling, there are a number of fit statistics that may be used to assess the fit of the item responses to the specified model. For Rasch modeling, by far the most commonly used fit statistics are Infit and Outfit (Wright & Stone, 1979). Both of these fit statistics are based on squared, standardized residuals. Infit and Outfit have two versions; a mean squares version and a standardized version. Infit and Outfit mean squares may be used to assess the magnitude of the misfit of an item, while the standardized versions may be used as a statistical significance test of the misfit.

Infit and Outfit, while useful fit statistics, have several flaws. First, Infit and Outfit are both measures of the difference in an item’s discrimination from the average discrimination of the remaining items on the test. They are not measures of the difference
between an item’s theoretical item response function (IRF) and the observed item response function (Wu & Adams, 2013). Thus, Infit and Outfit will be sensitive to any type of misfit that will affect an item’s discrimination. However, there are some types of misfit that Infit and Outfit will not be sensitive to. Additionally, standardized Infit and Outfit have exceedingly small type 1 error rates, far lower than the nominal value, for very easy or very difficult items. As a result, the power of standardized Infit and Outfit to detect measurement disturbances for these items may be reduced.

Because of their imperfections, it may be useful to supplement Infit and Outfit with an additional fit statistic. One such fit statistic, developed by Douglas and Cohen (2001), is known as the Root Integrated Squared Error (RISE). Unlike Infit and Outfit, RISE is a literal quantification of the differences between an item’s parametric and observed IRF. Therefore, it may be sensitive to types of misfit that Infit and Outfit are not sensitive to. When using RISE to compare the observed IRF to the theoretical IRF, the observed item responses may be smoothed in some way to create the observed IRF. There are several types of smoothing techniques that may be used, and there are varying amounts of smoothing that may be used within each technique.

The purpose of this dissertation is to compare the performance of RISE to standardized Infit and Outfit, in terms of type 1 error rates and power. Items with varying types of misfit will be generated to examine if RISE and standardized Infit and Outfit are differentially sensitive to certain types of misfit. Sample size will be varied to determine if there is a point at which all three fit statistics will have similarly high power, regardless of the type of misfit. Additionally, no consensus exists on an ideal type or amount of smoothing to use for calculating RISE. Thus, another purpose of this dissertation will be
to explore two types of smoothing techniques and nine amounts of smoothing within each technique.
CHAPTER 2
LITERATURE REVIEW

Item Response Theory

Item response theory (IRT) is a paradigm for describing what happens when a person encounters a test item. Specifically, use of an IRT model implies that the probability of a correct response from a person on an item is a function of a person’s latent trait, often symbolized by $\theta$, and one or more characteristics of the item. The relation between $\theta$ and the probability of a correct response on an item can be represented by an item response function (IRF), which depicts a monotonic, non-linear relation.

Multidimensional IRT exists, which allows for an item or a test to tap into multiple latent traits, but the focus of this dissertation will be on unidimensional IRT. Similarly, IRT can be used for items that are scored dichotomously or items that are scored using a rating scale. Dichotomous IRT will be the focus of this dissertation.

IRT is built upon two major assumptions: unidimensionality and local independence. Unidimensionality means that a single latent trait is driving responses to all of the items on the test. Local independence means that after conditioning on $\theta$, there is no relation between items on the test. This is an essential assumption, as estimation procedures for IRT rely on the local independence of items.

There are several models within unidimensional, dichotomous IRT, each named for the number of item parameters specified in the model. Birnbaum (1968) introduced several of these models. The 3 Parameter Logistic Model (3PL) is represented by the following equation:

$$
Pr \{ x_m = 1 \} = c_i + (1 - c_i) \frac{e^{\theta_i (b_i - h_i)}}{1 + e^{\theta_i (b_i - h_i)}},
$$

2.1
where \( \Pr\{x_{ni} = 1\} \) is the probability of correct response from person \( n \) on item \( i \), \( \theta_n \) is the ability of person \( n \), \( a_i \) is the discrimination of item \( i \), \( b_i \) is the difficulty of item \( i \), and \( c_i \) is a pseudo-guessing parameter. The \( a \)-parameter, conceptually, is a measure of how well an item can differentiate between persons of varying \( \theta \)’s. It is also the slope of the IRF at its point of inflection. The \( b \)-parameter is the item’s difficulty and can be defined as the \( \theta \) value at which the probability of a correct response is .5, after accounting for the pseudo-guessing parameter\(^1\). The \( c \)-parameter specifies a non-zero lower asymptote for the IRF, meaning that persons with extremely low \( \theta \)’s will have a non-zero probability of answering an item correctly. The \( c \)-parameter is used to model item responses when there is believed to be correct guessing in the data, as could be the case with multiple choice items.

If the \( c \)-parameter is constrained to zero, then the 3PL IRT model reduces to the 2PL IRT model, as shown in the following equation:

\[
\Pr\{x_{ni} = 1\} = \frac{e^{a_i(\theta_n-b_i)}}{1+e^{a_i(\theta_n-b_i)}},
\]

where the parameters are defined previously. While the 2PL model allows for items to have varying discriminations and difficulties across a test, it specifies that every item has a lower asymptote of zero. With the absence of a \( c \)-parameter, the \( b \)-parameter can be defined more simply as the \( \theta \) value at which the probability of a correct response is .5.

Finally, if the \( a \)-parameters for items on a test are constrained to equality, then the 2PL model reduces to the 1PL model, as shown:

\(^1\) Specifically, the \( b \)-parameter in a 3PL model is the \( \theta \) value at which the probability of a correct response is \((1 + c) / 2\), where \( c \) is the value of the \( c \)-parameter.
Pr \{x_{ni} = 1\} = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}, \tag{2.3}

with the parameters as defined previously. With the 1PL model, the only parametric
difference between items on a test is their difficulty. The 1PL model is mathematically
equivalent to the Rasch model, which will be the model of choice for this dissertation.

The Rasch Model

In the same decade that IRT was being formulated, Georg Rasch (1980/1960)
developed his own model for probabilistically describing encounters between persons and
items. As specified in the previous section, the Rasch model is mathematically equivalent
to the 1PL IRT model and can be represented by Equation 2.3. However, there are subtle
differences between the 1PL model and the Rasch model. These differences stem from
how the indeterminacy is solved during the estimation of parameters.

Given that \(\theta\) is a latent variable, it does not have an inherent metric. Thus, the
metric of \(\theta\) must be determined in some way. The distribution of \(\theta\) may be assumed to be
normal, but it does not have to be. Typically with the 1PL model, the variance of \(\theta\) is
specified to be 1. This allows for one common discrimination to be estimated across all
items. With the Rasch model, the discriminations of all items are set to 1. This allows for
the variance of \(\theta\) to be freely estimated. Therefore, when using the Rasch model, the
discrimination of items will affect the variance of \(\theta\), as opposed to affecting the value of
an \(a\)-parameter directly. Another difference between the 1PL model and the Rasch model
is how an origin is determined. With the 1PL model, the origin of the scale is specified to
be 0, and the mean item difficulty can be freely estimated. With the Rasch model, the
origin of the scale is set as the mean of the item difficulties, which allows for the mean of
\(\theta\) to be freely estimated.
**Beneficial properties of the Rasch model.** If the data fit the Rasch model, then the model possesses several desirable properties, some of which it shares with the 2PL and 3PL model and others which are unique to the Rasch model. A property that Rasch models and IRT models share is that of parameter invariance. This property specifies that item parameters calibrated using one sample of persons will equal the item parameters calibrated using a different sample of persons, up to a linear transformation. Similarly, person parameters will be invariant regardless of the items used, again up to a linear transformation. This is a crucial property that separates Rasch and IRT models from classical test theory, where item and person characteristics are sample dependent.

The Rasch model has several unique properties that separate it from the more complex 2PL and 3PL IRT models (Bond & Fox, 2013; Wright & Stone, 1977). One property is that the raw total scores for persons and items are the sufficient statistics for estimating person and item parameters. This means that all persons with the same raw total score will have the same estimated θ. Similarly, all items with the same raw total score will have the same estimated difficulty. This property only holds because the Rasch model specifies that all items have equal discriminations. With the 2PL and 3PL model, raw total scores are not sufficient statistics for parameter estimation. Because of this, parameters in the Rasch model can be stably estimated with fewer persons than when using the 2PL or 3PL model.

Another unique quality of the Rasch model is that of specific objectivity (Bond & Fox, 2013; Wright & Stone, 1977). The Rasch model places person abilities and item difficulties onto a common interval scale, specifically the logit scale. These estimates are objective in that the logit difference between two abilities does not vary across items and
the logit difference between item difficulties does not vary across persons. This allows for “person-free” comparisons of items and “item-free” comparisons of persons. This means that, for example, the difference in expected performance (expected log-odds) between two examinees on an item will be the same, regardless of which item is being examined.

**Importance of model-data fit with the Rasch model.** As described previously, the Rasch model specifies a parametric functional form for the probability of a correct response when a person encounters an item. Specifically, it is a function solely of the ability of the person and the difficulty of an item. This is a strict requirement. The desirable properties of the Rasch model, or any other IRT model, as explained in the previous section, only exist when the item responses fit the model. The item responses may not fit the Rasch model in numerous ways. For example, the presence of correct guessing in the data, which implies a non-zero lower asymptote for the IRF’s, would constitute a lack of fit of the data to the Rasch model.

A well known-property of the Rasch model is that item and person parameters may be severely biased if the item responses do not follow the functional form specified by the Rasch model. For example, calibrating 3PL items with the Rasch model will lead to underestimation of item difficulties, especially for the most difficult items or the least-able samples. In addition to parameter bias, properties like total score sufficiency and parameter invariance will not hold if the Rasch model is used to calibrate data that do not fit the Rasch model’s functional form. Thus, when using the Rasch model, it is vital to ensure that there is adequate fit of the data to the model. The fit of item responses to the Rasch model as a whole, or the fit of responses for a particular person or item, can be
summarized using fit statistics. In the next section, the concept of the residual, in the context of IRT and Rasch modeling, will be introduced. Most popularly used IRT/Rasch fit statistics will build upon the concept of the residual.

**Residuals in IRT/Rasch Modeling**

Generically, a residual in statistics is simply the difference between an observed value and a predicted/expected value. In the case of any dichotomous IRT model (including the Rasch model), an observed value is either a 0 for an incorrect response or a 1 for a correct response. The expected value is the probability of a correct response as implied by the specific IRT/Rasch model, which is a continuous variable that can range from 0 to 1. An IRT/Rasch model residual can formally be defined as:

$$y_{ni} = x_{ni} - p_{ni},$$

where $y_{ni}$ is the residual for person $n$ on item $i$, $x_{ni}$ is the observed response from person $n$ on item $i$, and $p_{ni}$ is the probability of a correct response from person $n$ on item $i$. The $p_{ni}$ is obtained by substituting the estimated $\theta$ for an item and estimated difficulty for an item into the specified IRT/Rasch model. There will be as many residuals as there are unique person/item encounters in the data set. The variance of an IRT/Rasch residual is the same as the variance of any dichotomy and can be expressed as:

$$w_{ni} = p_{ni}(1 - p_{ni}),$$

where $w_{ni}$ is the variance of the residual for person $n$ on item $i$ and $p_{ni}$ is defined as previously. The variance for an IRT/Rasch residual will be largest when $p_{ni}$ is .5 and will become increasingly smaller as $p_{ni}$ becomes smaller or larger than .5. Recall that the expected probability of a correct response will be .5 when an item’s difficulty is the same
as a person’s ability. This means that the variance of a residual will be largest when the item’s difficulty and person’s ability are well aligned. Positive residuals are associated with correct responses and negative residuals are associated with incorrect responses. Given that the predicted probability of a correct response will never be 0 or 1, an unstandardized IRT/Rasch residual will never be 0, even when the item response data fit the model.

There are two issues with the raw residuals as a diagnosis of data fit to an IRT/Rasch model. One issue is that each residual value has a different variance, and thus residuals are not directly comparable. A second issue is that the residuals for any one person or any one item will necessarily sum to zero. Thus, aggregating residuals across a person or item will not be informative when trying to diagnose misfit. To resolve the comparability issue, residuals can be standardized by dividing them by their respective standard deviations (Wright & Panchapakesan, 1969). This can be seen in the equation below:

$$z_{ni} = \frac{(x_{ni} - P_{ni})}{w_{ni}^{1/2}}$$

where $z_{ni}$ is the standardized residual for person $n$ on item $i$, the numerator is the raw residual defined previously and $w_{ni}^{1/2}$ is the standard deviation of the residual for person $n$ on item $i$. Standardized residuals are all on the same metric and can thus be directly compared. Smith (1988) has shown that Rasch standardized residuals are distributed approximately standard-normal across a variety of test conditions, such as test length and number of persons. These standardized residuals can be useful in detecting aberrant person/item encounters. However, standardized residuals still sum to zero across any one
person or any one item and cannot be aggregated to diagnose misfit. To remedy this issue, the standardized residuals can be squared, as shown below:

\[ z_{ni}^2 = \frac{(x_{ni} - p_{ni})^2}{w_{ni}}, \tag{2.7} \]

where all terms are defined as previously. Squared standardized residuals can both be compared directly to one another and can be aggregated across a person or an item as a diagnosis of that person’s or item’s fit to the IRT/Rasch model. In the following section, the most commonly used IRT fit statistics will be introduced, all of which build upon residuals.

**IRT Fit Statistics**

Instead of calculating a residual for each person \( n \), Yen (1981) introduced an IRT item fit statistic she called \( Q_1 \), which is a chi-squared statistic. Item fit statistics are not calculated for every person by item encounter, as with residuals. Instead, item fit statistics are calculated for each item. For Yen’s \( Q_1 \), examinees are first ordered according to their \( \theta \) estimates. Then, examinees are divided into 10 different groups, based on their \( \theta \) estimates. The formula for \( Q_1 \) is

\[
Q_1 = \sum_{k=1}^{10} \frac{N_k (x_{ik} - p_{ik})^2}{p_{ik} (1 - p_{ik})} \tag{2.8}
\]

where \( N_k \) is the number of examinees in ability group \( k \), \( x_{ik} \) is the observed proportion correct on item \( i \) for examinees in ability group \( k \), and \( p_{ik} \) is the mean predicted probability of a correct response on item \( i \) for ability group \( k \). Note that formula 2.8 looks very similar to the formula for a standardized residual in equation 2.7. The major difference is that for \( Q_1 \), students are grouped based on \( \theta \) before the residuals are calculated. Yen’s \( Q_1 \) is also very similar to a fit statistic developed by Bock (1972). One
difference was that Bock did not require a specific number of ability groups. The other difference is that the median of the \( \theta \) estimates within each ability group is used to calculate the predicted probability of a correct response for each ability group, not the mean as in \( Q_1 \). Yen suggested that the null distribution of \( Q_1 \) should approximate a chi-square distribution, although further research has shown that it does not.

McKinley and Mills (1985) developed an alternative to Yen’s \( Q_1 \) which they called \( G^2 \). Unlike \( Q_1 \), \( G^2 \) is a likelihood ratio statistic, not a chi-squared statistic. Although \( Q_1 \) and \( G^2 \) are asymptotically equivalent statistics, they may yield divergent results in practice. To compute \( G^2 \), examinees are first divided into 10 groups, according to their estimated \( \theta \)’s. Then, \( G^2 \) may be calculated using the following formula:

\[
G^2 = 2 \sum_{k=1}^{10} N_k \left[ x_{ik} \ln \left( \frac{x_{ik}}{P_{ik}} \right) + (1 - x_{ik}) \ln \left( \frac{1-x_{ik}}{1-P_{ik}} \right) \right]
\]

where \( \ln \) stands for the natural log and all other terms are defined as previously.

McKinley and Mills found that \( G^2 \) had acceptable type 1 error rates, more favorable than the fit statistics of Yen or Bock. They also found that, with large sample sizes (ie. 2,000), \( G^2 \) had good power when calibrating 3PL data with a 1PL model (\( > .8 \)), adequate power when calibrating 2PL data with a 1PL model (\( > .6 \)), and poor power when calibrating 3PL data with a 2PL model (\(< .3 \)). However, their study had only one replication, so conclusive statements regarding type 1 error and power rates are not warranted.

Orlando and Thissen (2000) discussed that Yen’s \( Q_1 \) and \( G^2 \) are not constructed like traditional chi-square goodness of fit statistics. This is because to obtain the model predicted probability of a correct response, examinees are grouped by their estimated \( \theta \). Therefore, examinees are grouped by an estimated latent variable, not an observed variable. Because \( Q_1 \) and \( G^2 \) may not follow a chi-square distribution, their type 1 error
rates and power may be affected. To combat this issue, Orlando and Thissen developed \( S-\chi^2 \) and \( S-G^2 \). These two fit statistics are calculated in ways that are almost identical to formulas 2.8 and 2.9. The primary difference is that examinees are grouped based on their raw total score, not an estimate of \( \theta \). Unlike a \( \theta \) estimate, raw scores are observed variables that are available before any IRT model is used.

As with the traditional \( Q_1 \) and \( G^2 \), the observed proportion correct can be found by simply dividing the number of examinees who got an item correct by the total number of examinees, within each raw score group. Calculating the expected proportion correct for each raw score group is much more complicated than when using \( Q_1 \) or \( G^2 \). Orlando and Thissen (2000) used a recursive algorithm alluded to by Lord and Wingersky (1984), which builds a joint likelihood distribution for each raw score group.

Orlando and Thissen (2000) compared the performance of the traditional \( Q_1 \) and \( G^2 \) to their newly developed \( S-\chi^2 \) and \( S-G^2 \) using a simulation study. They examined the type 1 error rates and power across varying test lengths (10, 40, or 80 items) and type of misfit (3PL data with a 1PL model, 2PL data with a 1PL model, and 3PL data with a 2PL model). Sample size was always 1,000 examinees. They found that \( Q_1 \) and \( G^2 \) both had unacceptably high type 1 error rates, especially when the length of the test was short. However, even when test length was long (80 items), both \( Q_1 \) and \( G^2 \) still performed poorly. The type 1 error rates for \( S-G^2 \) were better than that of \( Q_1 \) and \( G^2 \), but were still unacceptably large. Only \( S-\chi^2 \) displayed acceptable type 1 error rates across the varying test lengths.

Because inflated type 1 error rates can induce artificially inflated power, power could not be trusted for \( Q_1 \), \( G^2 \), and \( S-G^2 \). The power rates for \( S-\chi^2 \) ranged from .58 (3PL
data with a 1PL model) to .11 (3PL data with a 2PL model). Test length did not seem to have an effect on the power of $S$-$\chi^2$. Therefore, although the power of $S$-$\chi^2$ is not ideal, it does control for type 1 error rates and has become a popularly used fit statistic in IRT. For example, it is now the default fit statistic provided by Flexmirt (Cai & Wirth, 2013). While $S$-$\chi^2$ may certainly be used for assessing fit of items to the Rasch model, it is not the primary fit statistic used. In the next section, the fit statistics most commonly used in the Rasch model will be explored.

**Description of Infit and Outfit**

The two most popularly used fit statistics with Rasch modeling, Infit and Outfit, were first introduced by Wright and Stone (1979) for dichotomous Rasch models and Wright and Masters (1982) for polytomous Rasch models. Both of these fit statistics are residual based.

The Outfit mean square fit statistic is simply an average of the squared standardized residuals across a person or an item. This can be expressed as:

$$\text{Outfit} = \frac{1}{N} \sum \frac{(x_{ni} - P_{ni})^2}{w_{ni}},$$

where $N$ constitutes the number of persons if calculating Outfit for an item or the number of items if calculating Outfit for a person. Wright named this statistic “Outfit” because it is an outlier sensitive fit statistic, meaning it is sensitive to residuals from very unexpected responses. For example, a student of very high ability answering a very easy item incorrectly would constitute a highly unexpected response.

To reduce the impact of outlying residuals, Wright and Stone (1979) developed the Infit mean square fit statistic, which weights each squared standardized residual by its
statistical information. In the case of a Rasch residual, its information is its variance, which was defined in Equation 2.5. Infit can be defined as:

\[
\text{Infit} = \frac{\sum (x_{ni} - p_{ni})^2 w_{ni}}{\sum w_{ni}},
\]

where all terms are defined as previously. As shown, each squared standardized residual is multiplied by its variance. The larger the squared standardized residual, the smaller its variance, and thus the smaller its associated weight. Therefore, Infit is less affected by extreme outlying residuals than Outfit.

Both Infit and Outfit mean squares have an expected value of 1 (Wright & Stone, 1979). Values larger than 1 indicate that there is an underfit of the data to the Rasch model. Values smaller than 1 indicate overfit of the data, meaning the raw responses are too predictable or too Guttman-like. Wright and Linacre (1994) offered cut-off values of 0.8 to 1.2 for higher stakes tests and 0.7 to 1.3. for “run of the mill” tests. Thus, items with Infit and Outfit mean square values within that range would fit the model. However, as will be discussed in a subsequent section, no single appropriate critical value exists, in terms of statistical significance, for Infit and Outfit mean squares because their variances depend on the sample size.

To perform a statistical significance test of the fit of a person or item, Infit and Outfit mean squares can be standardized (Wright & Masters, 1982; Wright & Stone, 1979;). There are various procedures for standardizing Infit and Outfit mean squares, most of which are some type of cube-root transformation. Popular Rasch software Winsteps (Linacre, 2015) uses the Wilson-Haferty Transformation. The formula for standardizing Outfit mean square can be seen below:
\[ Z_{\text{Outfit}} = \frac{(Out^{1/3} - 1 + 2/(9N))}{(2/(9N))^{1/2}}, \]  

where \( Z_{\text{Outfit}} \) is the standardized version of Outfit and \( Out \) is the value for the Outfit mean squares obtained from Equation 2.10. Infit mean square can be standardized through an identical procedure. Standardized Infit and Outfit are purported to have an expected value of 0 and a variance of 1. Values above 0 indicate underfit of the data to the Rasch model and values below 0 indicate overfit. Values of -1.96 and 1.96 can be used as critical values for statistical significance tests using standardized versions of Infit and Outfit. As will be discussed in the subsequent section, studies have shown that standardized Infit and Outfit have fairly stable null distributions, thus allowing for the use of common critical values.

**Distributional Properties of Infit and Outfit**

Several studies have been conducted that relate to the distributional properties of Infit and Outfit, both the mean squares and standardized versions. General consensus exists concerning certain properties of Infit and Outfit, yet other properties have been met with inconsistent results. The designs and results of the various studies will be described in this section.

Smith (1991) was the first formal study concerning the distributions of Infit and Outfit, specifically the standardized versions. His study was a simulation consisting of mostly non-crossed factors of sample size, test length, item difficulty range, and alignment of test on examinee ability. Item responses were generated to fit the Rasch model. Thus, standardized Infit and Outfit results should have conformed to their expected null means and standard deviations of 0 and 1, respectively. Smith found that as the sample size increased, the mean of standardized Infit and Outfit departed further and
further from 0 in a negative direction. Thus, standardized Infit and Outfit values were increasingly underestimated as sample size increased. Standardized Infit exhibited noticeably more bias than standardized Outfit.

The effect of sample size on the bias of standardized Infit and Outfit means was moderated by test length, as the 20 item test exhibited less bias in standardized Infit and Outfit than the 10 item test. Sample size had an inconsistent effect on the standard deviations of standardized Infit and Outfit, although the standard deviations across sample sizes were generally less than the expected value of 1. As test length increased, both the means and standard deviations of standardized Infit and Outfit became less biased, with bias always being negative. As the range of item difficulties increased, the bias of the means of standardized Infit and Outfit increased (bias was again always negative) and the standard deviations followed an inconsistent pattern. As the alignment of the test on examinees worsened, the bias of the standardized Outfit mean worsened and the bias of standardized Infit was consistent (always negative bias). Test alignment did not have a consistent pattern of effect on the standard deviations of standardized Infit and Outfit.

Smith’s (1991) study suggested that the standardized versions of Infit and Outfit do not have stable null distributions across several factors. However, there were several limitations to his study. First, his factors were mostly non-crossed, so it was not possible to determine interactions between the factors. Second, each factor was replicated only 10 times, meaning a substantial amount of the differences in standardized Infit and Outfit means and standard deviations across levels of each factor may have been simply due to sampling error. Therefore, it is difficult to generalize the results beyond his study.
Smith and co-authors (1998) studied the distributional properties of the mean square versions of Infit and Outfit. They used a simulation study with two crossed factors of sample size (3 levels) and test length (2 levels), yielding six conditions in the study. Item responses were generated to fit the Rasch model so that the null distributions of Infit and Outfit mean squares could be studied. Each condition was replicated 100 times, as opposed to the 10 replications in Smith’s 1991 study, so the results should have been less affected by sampling error. Smith and co-authors found that the mean of the Infit and Outfit mean squares were stable around their expected value of 1, across sample size and test length.

The standard deviation of Infit and Outfit mean squares, however, changed across sample size, as one would expect for any effect size. Specifically, as sample size increased, the standard deviations of both Infit and Outfit mean squares decreased. Test length did not have an effect on the standard deviations of Infit and Outfit mean squares. Across all conditions, Infit mean squares had a smaller standard deviation than Outfit mean square. The lack of a stable null standard deviation for Infit and Outfit mean squares meant that a single critical value, such as 1.3, yielded different false hit-rates depending on the sample size. Specifically, false hit-rates decreased as sample size increased and were smaller for Infit than Outfit. Thus, use of a common critical value for assessing the statistical significance of Infit and Outfit mean squares would be inappropriate. In the Winsteps manual, Linacre (2015) made it clear the mean square versions of Infit and Outfit should not be used for assessing the statistical significance of an item’s misfit. Rather, Infit and Outfit mean squares should be viewed more as effect sizes, indicating the magnitude of the misfit. Thus, use of a cut-off value like 1.3 for Infit
and Outfit mean squares should be analogous to cut-off values like 0.2, 0.5, and 0.8 for Cohen’s $d$.

Smith and co-authors (1998) examined the type 1 error rates of the standardized Infit and Outfit across the six conditions. Recall that the standardized versions of Infit and Outfit are purported to follow a $z$ distribution and thus should be appropriate for assessing the statistical significance of item fit. They found that type 1 error rates for the standardized Infit and Outfit did decrease across levels of sample size. However, the effect of sample size on type 1 error rates for standardized Infit and Outfit was less pronounced than it was for the mean square versions of Infit and Outfit. As with the mean square versions, standardized Infit had smaller type 1 error rates than standardized Outfit. However, the type 1 error rates for both mean square and standardized versions were averaged across items, not reported for each individual item. Thus, the effect of item difficulties on type 1 error rates was not investigated.

Karabatsos (2000) critiqued the performance of both versions of Infit and Outfit, citing their faulty distributional properties. Karabatsos replicated the conditions used by Smith and co-authors (1998) and also found that type 1 error rates of Infit and Outfit mean squares varied depending on the sample size, thus precluding the use of a common critical value for assessing statistical significance. Karabatsos also found that the null mean of the standardized Infit and Outfit varied drastically with sample size. However, as highlighted by Wu and Adams (2013), the method in which Karabatsos increased the sample size was suspect. Karabatsos created 10 data sets, each with a different sample size. Instead of randomly drawing person and item parameters from a hypothetical distribution for each of the 10 data sets and simulating raw responses with those
parameters, Karabatsos simply took one set of item responses and duplicated that set of responses repeatedly to create data sets with increasingly larger sample sizes. This constitutes a clear violation of the assumption of independence and thus the results should not be trusted.

Wang and Chen (2005) conducted an item parameter recovery study for Winsteps. They simulated item responses to fit the Rasch model and created crossed factors of sample size (8 levels) and test length (4 levels), yielding 32 conditions. Each condition was replicated 500 times. Along with examining the recovery of item parameters, Wang and Chen also examined the distributional properties of both versions of Infit and Outfit. They found that the mean of Infit and Outfit mean squares were almost exactly 1 across all conditions. The standard deviations, as should be expected, decreased as sample size increased. Across all conditions, the standard deviation of Infit mean squares was smaller than that of Outfit mean squares. This aligned with the findings of Smith (1998) and Karabatsos (2000).

The standardized versions of Infit and Outfit had means just slightly below 0 across all conditions, which stands in contrast to the findings of Smith and Karabatsos. Averaged across items, the standard deviations of standardized Infit and Outfit were less than the expected value of 1, but not drastically so (ranging from 0.79 to 0.97). Wang and Chen also examined the standard deviations of standardized Infit and Outfit at the item level. They found that items that were well targeted to the persons had standard deviations very close to 1 for both standardized Infit and Outfit. As items became increasingly poorly targeted in either direction of difficulty, the standard deviations of standardized Infit and Outfit became much smaller. This was especially true for
standardized Infit. Thus, Wang and Chen found that standardized Infit and Outfit had stable null distributions for well targeted items and distributions that were too narrow for poorly targeted items, thus leading to type 1 error rates that were too small for poorly targeted items.

Wu and Adams (2013) studied and described the distributional properties of the mean square and standardized version of Outfit. In alignment with previous studies (Karabatsos, 2000; Smith, 1998; Wang & Chen, 2005), Wu and Adams demonstrated that while the Outfit mean square maintained its expected value of 1 across sample sizes, the standard deviation of Outfit mean square was noticeably affected by sample size, an expected property for an effect-size measure. Specifically, the variability of Outfit mean squares for 20 items simulated to fit the Rasch model was much smaller when sample size was 800 as opposed to when sample size was 100. Wu and Adams agreed with previous studies in that a single common critical value for Outfit mean square would be inappropriate, if one wanted to use the mean square versions as a statistical significance test.

Wu and Adams (2013) also discussed the contrary findings of Karabatsos (2000) and Wang and Chen (2005) in terms of the distribution of standardized Outfit. Wu and Adams replicated Karabatsos’s analysis and found results similar to his. However, as discussed previously, they noted the inappropriateness of his data generation. When data were generated appropriately, they found that standardized Outfit had its expected mean of 0 and standard deviation of 1, in support of the results found by Wang and Chen. They noted that the power of standardized Outfit will of course increase as sample size
increases, as does power with virtually any statistical significance test, but the null
distribution of standardized Outfit should be fairly stable.

DeMars (2017) conducted a simulation study to explore the use of mean square
and standardized versions of Infit and Outfit concurrently, with standardized versions
used for statistical significance and mean square versions to assess the magnitude of
misfit. Responses to 40 items were generated to fit the Rasch model, along with six
misfitting items. Sample sizes of 100 and 5,000 were used, and each condition was
replicated 3,000 times. DeMars found that the mean of Infit mean square was nearly 1 for
all fitting items, regardless of the sample size. The mean for Outfit mean square slightly
deviated from 1, but this was not considered practically significant. The standard
deviation of Infit and Outfit mean squares were smaller when the sample size was large,
confirming findings of previous studies. Confirming Wang and Chen (2005), DeMars
found that standard versions of Infit and Outfit had standard deviations near 1 for well-
targeted items, but standard deviations that were too small for poorly targeted items.
Averaged across items, this led to type 1 error rates that were lower than the nominal
value of .05, especially for Infit. This was due to the presence of items difficulties that
were far from the mean person ability. As expected, the power of standardized Infit and
Outfit to detect the misfitting items increased drastically as sample size increased from
100 to 5,000.

In summary, numerous studies have investigated the distributional properties of
both versions of Infit and Outfit. There is universal agreement that while the mean square
versions of Infit and Outfit have a stable null mean of 1, their standard deviations
significantly decrease as sample size increases. This is a typical property of effect sizes.
This lack of a stable null distribution means that no single cut-off value can be used for Infit and Outfit mean squares to assess statistical significance. In terms of the standardized versions of Infit and Outfit, there has been discrepancies in the results across studies. However, the more recently and robustly conducted studies (DeMars, 2017; Wang & Chen, 2005; Wu & Adams, 2013) have demonstrated that standardized Infit and Outfit have fairly stable null distributions for well-targeted items, and distributions that are too narrow for poorly targeted items.

**Limitations of Infit and Outfit**

It is important to understand the limitations of Infit and Outfit, as well as their appropriate interpretations. As highlighted in the previous section, mean square versions of Infit and Outfit may be used as an effect size. However, due to the lack of a common critical value across sample sizes, mean square versions of Infit and Outfit should not be used to assess the statistical significance of item or person fit. The standardized versions of Infit and Outfit can be used to assess statistical significance, as long as the user is aware that both statistics will be too conservative with poorly targeted items.

When using Infit and Outfit to assess item fit, it is also crucial to interpret the statistics appropriately. As highlighted by Wu and Adams (2013), Infit and Outfit are not a measure of the difference between the observed IRF and the Rasch implied IRF. Instead, Infit and Outfit are measures of whether the discrimination of an item is different than the average discrimination of the other items on the test. Infit and Outfit will be relatively sensitive to any type of misfit of the data to the Rasch model that affects the discrimination of the item in question, relative to the other items on the test. For example, an item that follows a 2PL model with an \( a \)-parameter of 0.5 will be flagged by Infit and
Outfit as underfitting the model, if the majority of the remaining items follow the Rasch model. Similarly, an item that has a non-zero lower asymptote will also be flagged as misfitting, given that a non-zero lower asymptote can be conceptualized as an extreme lack of discrimination at the low end of the ability scale.

Many types of misfit may lead to a change in the discrimination of an item, relative to the other items on a test. Thus, Infit and Outfit will be sensitive to many types of misfit. However, there are examples of misfit that do not change the relative discrimination of an item and may not be identified by Infit and Outfit. For example, examine the two IRFs in Figure 1. The solid line is the Rasch implied IRF, whereas the dotted line is the observed IRF. Through a simple visual inspection, the observed IRF appears markedly different than the Rasch implied IRF. One would assume that Infit and Outfit would identify this item as misfitting, given that the observed data clearly do not conform to the Rasch model. However, the discrimination of the observed IRF is nearly identical to the discrimination of the Rasch implied IRF. Therefore, Infit and Outfit would not flag this item as misfitting, even though it clearly does not follow the Rasch model. Given this limitation of Infit and Outfit, it may be useful to supplement Infit and Outfit with a fit statistic that is a quantification of the difference between the observed and Rasch implied IRFs.

**Smoothing**

In subsequent sections of this literature review, a fit statistic called the Root Integrated Squared Error (RISE) will be introduced and discussed. One aspect of RISE is that it requires smoothing of the observed data to create an observed IRF. Therefore, this
section of the literature will focus on smoothing and several methods by which it can be performed.

Smoothing, in a statistical sense, is a process for adjusting observed data so that it approximates a smooth function. The purpose of smoothing is to remove “roughness” in the data that may simply be due to error, while still capturing important patterns in the data. Smoothing can be applied to myriad types of data, including dichotomously scored item responses. For example, raw responses to an item may be smoothed using any number of smoothing techniques, creating a non-parametric IRF that depicts the relation between the probability of a correct response and some measure of student ability, such as the total score on the test. This stands in contrast to parametric IRT/Rasch modeling. With parametric IRT/Rasch modeling, it is assumed that the relation between the probability of a correct response on an item and examinee ability can be completely defined by a small set of parameters. In contrast, using smoothing procedures to create an IRF requires no assumptions other than that the relation between probability of a correct response and ability can be represented by some type of smooth curve. This allows for the data to be freed from a “parametric straightjacket” (Simonoff, 2012) when conducting statistical analyses such as modeling item responses.

There are numerous methods that can be used to smooth data. For brevity, only smoothing methods used for this dissertation will be discussed. One such smoothing method, known as Hanning, can be attributed to Julius Von Hann (1903), a 20th-century climatologist. Vellemen and Hoaglin (1981) provide an excellent formal description of the Hanning procedure. Hanning uses running weighted averages to smooth raw data and can be represented by a simple formula:
where \( z_i \) is a smoothed value for data point \( i \), \( y_i \) is the data point being smoothed, and \( y_{i-1} \) and \( y_{i+1} \) are the data points immediately before and after the data point being smoothed, respectively. This process is carried out for every data point in the data set. The smoothed values from the Hanning technique may then be re-smoothed as many times as deemed necessary to create a smoothed curve. The weights can take on any value, as long as the weights sum to 1. Inherent to the Hanning technique is that each data point is smoothed using only three data points: the data point itself and the two data points immediately before and after it. In successive iterations, points further away contribute indirectly to point \( i \) through their impact on points \( i-1 \) and \( i+1 \) in previous iterations.

In the case of item response data, the Hanning technique could be conducted in several ways. For example, consider item response data represented by a simple two-dimensional graph. The \( x \)-axis of a graph could represent examinees, arranged by some measure of ability, from least able to most able. The \( y \)-axis would constitute dichotomously scored responses, either 0 or 1. These raw responses could be smoothed using the Hanning technique, with the smoothing conducted many times to create a smooth observed IRF. Alternatively, one could first group examinees by some measure of ability and then use the Hanning technique to smooth the observed proportion correct at each ability group. Grouping examinees by an ability measure first, before smoothing, would require less amounts of re-smoothing, given that grouping examinees is in itself a crude form of data smoothing.

As stated earlier, Hanning is a conceptually simple smoothing technique in that each data point is smoothed using only three values. A more complex, but commonly
used smoothing technique in statistics is kernel smoothing. Ramsey (1991) introduced the use of kernel smoothing to smooth item response data. Instead of smoothing a data point by using only the data point itself and the data points immediately before and after, as with Hanning, kernel smoothing can be used to assign smoothing weights to all data points in a data set, shown generically as:

\[ z_i = w_{i-1}y_{i-1} + w_iy_i + w_{i+1}y_{i+1} + \ldots + w_ny_n, \]  

2.14

where weights \( w \) are assigned to every value in the data set, from \( y_1 \) to \( y_n \), the data points farthest from \( y_i \) on each side of \( y_i \). The relative weights \( w \) are determined by use of a kernel function \( K(u) \):

\[ K\left( \frac{q_i - y_i}{h} \right), \]  

2.15

where \( q_i \) is the distance in data points from \( y_i \) so that the weights are maximized when \( u = 0 \) and fall to zero on either side of \( y_i \). The character \( h \) represents a bandwidth parameter and is used to control how quickly the weights reduce to zero on either side of \( y_i \). Smaller bandwidth parameters will be associated with more rapidly decreasing weights. There are several choices for kernel functions. One commonly used kernel function when smoothing item response data is the Gaussian (Normal) kernel function, which is represented by the following formula:

\[ K(u) = \exp(-u^2 / 2), \]  

2.16

When a Gaussian kernel function is used to define smoothing weights, weights will be largest for data points near \( y_i \), and will decrease in a normally distributed fashion on either side of \( y_i \). Again, the bandwidth parameter \( h \) will define the shape of this normal
distribution of weights, with smaller bandwidths associated with a tighter distribution of weights and thus a more dramatic decrease in weights on either side of $y_i$.

**The Root Integrated Squared Error**

Infit and Outfit, although they are the most popularly used, are not the only methods of assessing the fit of data to the Rasch model. One common and simple way of assessing fit in any item response model is to graphically compare the theoretical IRF to the observed IRF. Georg Rasch (1980/1960), in his development of the Rasch model, was interested in using graphical means of assessing the fit of item response data to his model. However, the issue with simply comparing the theoretical IRF to the observed IRF is that it is a subjective comparison. Two researchers examining the fit of an item could arrive at very different conclusions when making the same graphical comparison. Georg Rasch did not have a formal method of quantifying the difference between the theoretical Rasch IRF and the observed IRF.

Douglas and Cohen (2001) developed a method for quantifying the difference between a theoretical IRF and an observed IRF, which they called the Root Integrated Squared Error (RISE). RISE can be represented by the following formula:

$$RISE = d(P, P^*) = \left[ \int \left( P(\theta) - P^*(\theta) \right)^2 f(\theta) d\theta \right]^{1/2}, \tag{2.17}$$

where $P(\theta)$ is the theoretical IRF, $P^*(\theta)$ is the observed IRF, and $f(\theta)$ is the density of $\theta$. Given that calculating RISE requires formal integration, a discrete approximation of RISE was alluded to by Douglas and Cohen and formally addressed in Wells and Bolt (2008). This discrete approximation takes the form:
where $\hat{P}_{ji}$ is the parametric IRF, $\hat{P}_{non,ji}$ is the non-parametric (observed IRF), and $Q$ is the number of evaluation points across the IRFs. Thus, RISE is a literal quantification of the difference between the theoretical and observed IRF. When calculating RISE, Douglas and Cohen weighted the numerator by the density of $\theta$ at each evaluation point, as indicated by the $f(\theta_j)$ term. However, Wells and Bolt weighted all points equally.

The observed IRF is derived non-parametrically, whereas the theoretical IRF is based on a specified parametric model that states that the probability of a correct response on an item can be represented solely by a small number of parameters. Thus, the observed IRF is based on fewer assumptions than the theoretical IRF. If there are discrepancies between the observed IRF and theoretical IRF, than the observed IRF can be considered a more accurate reflection of the “true” IRF. Thus, departure of the theoretical IRF from the observed IRF would constitute a lack of fit of the parametric model to the data. RISE is a quantification of this departure.

To calculate RISE, some type of statistical smoothing technique can be used to create the non-parametric IRF. Douglas and Cohen (2001) proposed the use of kernel smoothing with a Gaussian kernel function and used bandwidth parameters of 0.2 and 0.5 for their first and second simulated demonstrations, respectively. As discussed previously, use of a Gaussian kernel function applies weights that are largest around the data point being smoothed, which in this case is the proportion correct on an item for students at a given $\theta$ value. Weights decline on either side of $\theta$.  

\[
RISE_i = \sqrt{\frac{\sum_{j=1}^{Q} (\hat{P}_{ji} - \hat{P}_{non,ji})^2 f(\theta_j)}{Q}},
\]

2.18
For truly non-parametric IRF estimation, however, examinees’ θ’s are unknown and thus cannot be used to specify smoothing weights. Douglas and Cohen proposed a solution to this issue. First, each persons’ percentile on the total score distribution is calculated, but not including the item being assessed. Then, each persons’ percentile is converted to the θ value associated with that percentile in whatever θ distribution corresponds to the IRF. For example, when using Marginal Maximum Likelihood for parametric IRF estimation, the distribution may be estimated along with the item parameters, or it may be specified as standard normal. For example, in a standard normal distribution, as Douglas and Cohen point out, a person with a raw score at the 95th percentile would be assigned a θ of 1.645 for the estimation of the non-parametric IRF. After converting all persons’ percentiles into θ’s, the proportion correct at each θ may be smoothed using the kernel smoothing procedure.

Once the observed data has been smoothed to create a non-parametric IRF, it can be compared to the parametric IRF and differences between the two IRFs can be quantified using RISE. However, the magnitude of RISE is not easily interpreted, as RISE does not follow a known sampling distribution, unlike the standardized versions of Infit and Outfit. Thus, a statistical significance test for RISE cannot be conducted simply using a distribution like $\chi^2$ or $z$. Instead, statistical significance tests of RISE can be conducted by using a bootstrapping procedure to create an empirical sampling distribution of RISE.

First, estimate item parameters using the specified model, which was the 2PL IRT model in the case of Douglas and Cohen (2001). Next, randomly generate $n$ θ’s from a specified distribution, with $n$ representing the sample size. Then, generate item responses
for every person/item encounter using the simulated $\theta$’s and the estimated item parameters, using the specified model. Next, group examinees based on some measure of ability, like the raw total score. Smooth the proportion correct values at each ability level to create non-parametric IRFs for each item. Estimate the parametric IRF for each item. Then, calculate RISE for every item using the non-parametric IRF, the parametric IRF, and Equation 2.16. Repeat this process 499 more times for a total of 500 replications. Using this process, an empirical sampling distribution of RISE has been created for each item of interest. This empirical sampling distribution may be conceptualized as the null distribution of RISE, as responses to each item in each replication were generated to fit the specified model. If a .05 alpha level is desired for the statistical significance test, find the 95th percentile of the empirical RISE distribution for each item to determine the critical value for each item. If an item’s RISE value calculated using the real data exceeds the RISE critical value from the empirical sampling distribution for that item, then the item has statistically significant misfit.

**Studies Investigating RISE**

Although Douglas and Cohen (2001) proposed RISE and conducted several simulations concerning RISE, the simulations contained only one replication and were intended for illustrative purposes only. Wells and Bolt (2008) were the first to formally study the properties of RISE, in terms of Type 1 error rates and power. They conducted a simulation study to compare the Type 1 error rates and power of four IRT fit statistics: $G^2$, $S$-$X^2$, RISE, and RISE*. For RISE, smoothing was conducted using a uniform kernel function, whereas a Gaussian kernel function was used for RISE*. In each replication of the simulation study, they simulated responses to a set of operational items and to a set of
pilot items. The operational items were used to set the metric of the θ scale and the pilot 
items were used to assess Type 1 error rates and power of the various fit statistics.

The simulation study contained three crossed factors: number of operational 
items (10, 20, 40, or 80 items), sample size (250, 500, or 1000 examinees) and percentage 
of misfitting operational items (0%, 10%, 30%, or 50%). The number of pilot items was 
always 40. Responses to both the operational and pilot items without misfit were 
generated according to the 2PL model. Responses to misfitting operational and pilot 
items followed one of three models: the 3PL model, the mixture nominal response model, 
or the hyperbolic cosine model. Thirty-two of the pilot items were always misfitting 
items, regardless of the condition. These were used to assess the power of the fit 
statistics. Thus, 8 of the pilot items in each condition fit the 2PL model and were used to 
assess type 1 error rates. The number of misfitting items on the operational test depended 
on the condition, as specified above. Each condition was replicated 1,000 times.

Wells and Bolt (2008) found that the presence of misfitting operational items did 
not affect type 1 error rates or power when evaluating the pilot items, so results were 
averaged across levels of that factor. They found that $S\cdot X^2$, RISE, and RISE* had type 1 
error rates near the nominal level of .05, regardless of sample size and test length. $G^2$, 
however, had type 1 error rates greater than the nominal alpha for short tests. In terms of 
power, RISE and especially RISE* performed more favorably than did $S\cdot X^2$ or $G^2$. This 
was the case regardless of test length and sample size.

Liang and Wells (2009) conducted a similar study of RISE, but applied RISE to 
detecting misfit of polytomous items, specifically the generalized partial credit model. 
Liang and Wells (2015) also studied the use of RISE for assessing item fit on a mixed
format test, meaning a test with both dichotomous and polytomous items. In both studies, Liang and Wells found that RISE outperformed more popularly used fit indices, in terms of type 1 error rates and power. Thus, through the study of Wells and Bolt and the studies of Liang and Wells, there is evidence that RISE may be an effective statistic for evaluating the fit of data to item response models.

Studies mentioned previously all examined RISE in the context of IRT models with varying item difficulties and discriminations, be it dichotomous or polytomous models. Jennings and Engelhard (2017) were the first to formally investigate the performance of RISE to evaluate the fit of item response data to the Rasch model. Specifically, they compared the performance of RISE to commonly used Rasch fit statistics Infit and Outfit, in an applied setting. They used data consisting of responses to 25 statistics items from 1,255 undergraduate students. They found that RISE and standardized Infit and Outfit did not flag the same items as misfitting. Specifically, standardized versions of Infit and Outfit flagged many more items as misfitting than did RISE. However, given that this was an applied study, it is not possible to assess whether RISE or Infit and Outfit were more accurate, just that they performed differently.

The Jennings and Engelhard (2017) study differed significantly from previous studies. Although they followed the bootstrapping procedure laid out by Douglas and Cohen (2001) for assessing the statistical significance of RISE, their method of smoothing differed from the studies mentioned previously. Jennings and Engelhard did not use kernel smoothing to create the observed IRFs. Instead, they used Hanning, with the specific weights shown in Equation 2.11. If they had used kernel smoothing, the performance of RISE may have differed. Furthermore, instead of grouping examinees by
some measure of ability before smoothing, like total score, they conducted smoothing on
the individual raw item responses. The number of smoothing iterations used for each item
depended on the number correct responses for each item. For example, if 50 examinees
answered an item correctly, that item would undergo 50 iterations of Hanning smoothing.
They did not provide justification for this decision, although they wrote that one could
expect the amount of smoothing to affect results and that subsequent studies should
examine the impact of various amounts of smoothing. If persons had been grouped by
their estimated θ first and then smoothed, the results likely would have varied as well.

An additional difference between the study of Jennings and Engelhard and the
work done by Douglas and Cohen and Wells and Bolt is that Jennings and Engelhard
used the IRT-based θ when estimating the \( \hat{P}_{ji} \) term (the expected IRF) in Equation 2.18.
In contrast, the other studies used a transformation of the rest score\(^2\). With the 2PL
model, any given rest score corresponds to multiple IRT θ estimates. Therefore, \( \hat{P}_{ji} \) could
not be calculated simply by using the parameters from a conventional IRT calibration.
Instead, the 2PL parameters were a function of the transformed rest score and were
estimated by finding the parameters which provided the closest fit to the non-parametric
IRF. In contrast, because Jennings and Engelhard used the Rasch model, there was a one-
to-one correspondence between the total score and the estimated Rasch ability. In
estimating \( \hat{P}_{ji} \) they simply calibrated the data using Winsteps and used the model
predicted \( \hat{P}_{ji} \). In addition to simplicity, this has the advantage of comparing the non-
parametric IRF directly to the parameters obtained through standard calibrations.

\(^2\) The rest-score is the raw total score on the test, after removing the item in question
Research Questions

There is previous literature concerning the performance of RISE, in comparison to commonly used fit statistics in terms of type 1 error rates and power. However, although smoothing techniques were explored in previous studies, they were not a major focus of the studies. Further, only one study exists concerning the comparison of RISE to popular Rasch fit statistics Infit and Outfit. This study was an applied study, so conclusive statements about the performance of RISE compared to Infit and Outfit could not be made. Therefore, this dissertation will address several new research questions.

Research Question 1: How does sample size affect the type 1 error rates and power of RISE and standardized versions of Infit and Outfit?

Given all three of these statistics can be assessed for statistical significance, sample size is of course going to affect power, whereas it should not affect type 1 error rates. Previous studies have shown that, as expected, the power of all three statistics increases as sample size increases.

Research Question 2: How does the amount of smoothing affect type 1 error rates and power of RISE?

Several of the studies related to RISE mention the potential impact that the amount of smoothing may have on RISE. Under-smoothing can overemphasize roughness in the data that is due to chance or error, whereas over-smoothing can obscure meaningful departures of the observed IRF from the theoretical IRF. Ideally, a middle ground would be determined that does not under-smooth or over-smooth the data. However, there is no consensus on how much smoothing is appropriate. For example, Douglas and Cohen (2001) wrote that “it is common for researchers to select a bandwidth that results in an
estimated function that achieves some minimal level of smoothness. This is determined by visually inspecting the estimated function” (Douglas & Cohen, 2001 p. 236). While visual inspections can be informative, this research question concerning smoothing amount will be empirically assessed to determine whether a “happy medium” amount of smoothing really is most appropriate. Nine amounts of smoothing will be investigated for each type of smoothing.

Research Question 3: Does the type of smoothing technique affect type 1 error rates and power of RISE for the Rasch model?

Smoothing is a vital step in creating the observed IRF that is compared to the theoretical IRF when calculating RISE. Other than Wells and Bolt (2008) comparing Gaussian and uniform kernel smoothing, no research exists formally comparing the impact of smoothing technique on performance of RISE. Wells and Bolt found that the results differed, depending on the type of smoothing being used. Research question 3 will allow for further examination of the effects of smoothing on RISE. Two smoothing techniques will be investigated: kernel smoothing with a Gaussian function and Hanning with examinees first grouped by ability. For the Hanning procedure, the amount of smoothing will be defined by the number of smoothing iterations. For kernel smoothing, the amount of smoothing will be defined by the bandwidth parameter. Kernel smoothing with a uniform function was not chosen for two reasons. One is that RISE performed better with a Gaussian kernel. Secondly, there was simply not enough time, in terms of simulation time, to include a third smoothing technique.

Research Question 4: Does RISE have different power rates than standardized Infit and Outfit across various types of item misfit?
RISE is a quantification of the observed IRF and the theoretical IRF, whereas Infit and Outfit essentially measure whether the discrimination of an item is different than the average discrimination of the rest of the items. Given that RISE and Infit/Outfit are computed and interpreted differently, it is reasonable to expect that they might have different power rates as well, especially with certain types of items. Seven types of misfitting items, which will be described in the method section, will be used to explore differences in power between RISE and standardized Infit and Outfit.
CHAPTER 3

METHODOLOGY

Overview

The research questions in this study cannot be addressed using real data. With applied data, one can determine which null hypotheses to reject or fail to reject, but it is unknown whether those decisions are correct or incorrect. Therefore, the data used in this dissertation were simulated. The use of simulated data allows a researcher to specify the exact properties of the data, allowing “truth” to be known. In addition, the use of simulated data allows for conducting many replications of data creation and analyses. These replications can then be used to examine statistical properties like type 1 error rates and power, both of which are of interest in this dissertation. In this chapter, the various factors of this simulation study will be described in detail, along with how the data were generated to conform to these factors.

Design

Three factors were manipulated in this study: smoothing method, smoothing amount, and sample size. The two smoothing methods used in this study were Hanning when examinees were grouped by their estimated θ and kernel smoothing with a Gaussian function. Nine amounts of smoothing were used for each smoothing method to represent a wide range of possible smoothing amounts. Because the specific amount of smoothing differed across the two smoothing methods, smoothing amount was nested within the smoothing method. Sample size was either 200, 500, or 1,000 examinees. The sample size factor was crossed with the other two factors of the study. This yielded 54 overall conditions in the study. Each condition was replicated 500 times to ensure stable
estimates of type 1 error and power for each condition. Regardless of the levels of the factors, examinees always responded to the same 50 item test. The method by which item responses were generated, along with more detail concerning each manipulated factor, will be described in subsequent sections.

**Simulating Item Responses**

Responses to a 50-item test were generated for 200, 500, or 1,000 examinees, depending on the level of the sample size factor. Regardless of the condition, examinee \( \theta \)'s were drawn from a standard normal distribution. Responses to the first 43 items on the test were generated using the Rasch model, as specified in Equation 2.3. Item difficulties for these 43 items ranged from -2.1 to 2.1 logits, in increments of 0.1. Thus, they were uniformly distributed with a mean of 0. These items may be considered the operational items of the test that have been quality controlled to ensure that they fit the model. The remaining seven items (item 44 through item 50) may be considered pilot items. Responses to each of these items were generated to represent a specific type of item misfit.

Responses to item 44, which can be called the wavy item, were generated using the following function:

\[
\Pr \{x_{ni} = 1\} = 3PL + 0.8(0.5 - abs(0.5 - 3pl)) \sin(1.5(\theta_n - b)) \pi ,
\]

where \( 3PL \) represents the probability obtained from Equation 2.1, \( abs \) indicates absolute value, and \( \pi \) is the commonly used constant. For this item, the \( a \)-parameter was set to 1, the \( b \)-parameter was 1, and the \( c \)-parameter was 0. A visual depiction of this function is provided in Figure 1, which was used in Chapter 2 when describing the interpretation of
Infit and Outfit. As can be seen, the IRF for item 44 is non-monotonic, with probabilities of a correct response fluctuating wildly across the θ continuum.

Responses to item 45, which can be called the big dip item (Van Rijn et al., 2016), were generated using the following function:

\[
\Pr\{x_m = 1\} = \frac{c}{1 + e^{a(\theta_i - b_h + d_h)}},
\]

where \(d_i\) distorts the IRF such that the probability of correct response dips in the middle of the ability range, and all other terms are defined as previously. The \(a\)-parameter was 4.25, the \(b\)-parameter was 1, and the \(c\)-parameter was .2, and the \(d\)-parameter was 1.5.

Figure 2 provides a visual depiction of item 45, as compared to the best-fitting Rasch model IRF. As the name of the item implies, there is a large decrease in the probability of correct response in the middle of the IRF. This item is highly discriminating in some areas of the θ continuum (i.e., 0 to 2) and completely lacking discrimination in the low end of the θ continuum due to the presence of a non-zero lower asymptote.

Responses to item 46 were generated using the 3PL model, as specified in Equation 2.1. The \(a\)-parameter was 1, the \(b\)-parameter was 1, and the \(c\)-parameter was .2. The item was generated to be difficult so that the non-zero lower asymptote was consequential. If an item is very easy relative to the examinee population, even if the item has a true non-zero lower asymptote, its presence will not affect calibration of the item with the Rasch model because there are so few people located on the θ scale where the asymptote manifests itself. Figure 3 is a visual depiction of the IRF for item 46, clearly displaying its non-zero lower asymptote.

Responses to item 47 were generated using the 2PL model, as specified in Equation 2.2. The \(a\)-parameter was 2 and the \(b\)-parameter was 0. Given the value of the
$a$-parameter, item 47 was much more discriminating than the 43 operational items, all of which had a discrimination of 1. Figure 4 displays the IRF for item 47. It may seem counterintuitive to describe item 47 as misfitting, given that high discrimination is a desirable quality of an item. Of course, a practitioner would want all items to behave like item 47. However, item 47 would constitute misfit to the Rasch model, given that the 43 operational items all have discriminations of 1. The Rasch model specifies that all items have equal discriminations. If that requirement is violated, desirable properties like sufficiency of number-correct for estimation and specific objectivity are no longer tenable. Thus, in this study, item 47 would violate the Rasch model, with its responses being deemed as “overfitting” or “too Guttman-like.”

Responses to item 48 were generated using the 4PL model, which is represented using the following function:

$$\Pr \left\{ x_m = 1 \right\} = c_i + (u_i - c_i) \frac{e^{a_i(\theta_i - b_i)}}{1 + e^{a_i(\theta_i - b_i)}},$$

where $u_i$ represents the upper asymptote of the IRF and all other terms are defined as previously. For item 48, the $a$-parameter was set to 3, the $b$-parameter was -1, the $c$-parameter was 0, and the $u$-parameter was .8. Figure 5 depicts the IRF for item 48. As can be seen, item 48 is highly discriminating across much of the $\theta$ continuum. However, for high ability examinees, the item is completely undiscriminating, as indicated by the non-1 upper asymptote.

Responses to item 49 were generated using the hyperbolic cosine model (Wells & Bolt, 2008), represented by the following function:

$$\Pr \left\{ x_m = 1 \right\} = \frac{e^{b_i}}{e^{b_i} + 2 \cosh(\theta_n - 0.75)},$$

where $b_i$ represents the upper asymptote of the IRF and all other terms are defined as previously.
where the cosh term represents the hyperbolic cosine and all other terms are defined as previously. Figure 6 depicts the IRF of item 49. As can be seen, the hyperbolic cosine IRF aligns with the Rasch IRF for much of the θ scale. However, as θ increases into the upper end of the θ scale, the probability of a correct response actually decreases, with examinees of very high ability associated with lower probabilities of a correct response.

Finally, responses to item 50, which can be called the flat middle item, were generated using the following function:

\[
\Pr\{x_{ni} = 1\} = 0.55 \left( \frac{e^{5.95(\theta_n + 1)}}{1 + e^{5.95(\theta_n + 1)}} \right) + 0.45 \left( \frac{e^{5.95(\theta_n - 2)}}{1 + e^{5.95(\theta_n - 2)}} \right)
\]

3.5

The function in 3.5 is a mixture model, with the values of 0.55 and 0.45 constituting the weights. Figure 7 displays the IRF for item 50. As the name implies, item 50’s IRF is highly discriminating, except for the flatness in the middle of the θ scale.

Responses to the fifty items were calibrated using Winsteps (Linacre, 2015). To keep the misfitting items from influencing the parameter estimates and fit statistics of any other item, these items were not included in the calibration of θ in the joint calibration. In Winsteps, this can be done automatically using the IWEIGHT command. The forty-three operational items were given weights of 1, which is the default in parameter estimation. The seven misfitting pilot items were given weights of 0. This means that parameter estimates and measures of fit were provided for these items, but the items did not influence the parameter estimates or fit statistics of any other item on the test. This allowed for the examining of type 1 error rates and power simultaneously, which will be discussed in more detail in subsequent sections.
The traditional Rasch fit statistics, standardized Infit and Outfit, were obtained directly from the Winsteps software. For the remainder of this dissertation, only the standardized versions of Infit and Outfit will be explored and discussed. Thus, any mention of Infit and Outfit refers to the standardized versions. The statistical significance of both Infit and Outfit were recorded for each item in each replication of the study. Critical values of -1.96 and 1.96 were chosen, in order to align with an alpha level of .05.

**Calculating RISE**

It is very important to distinguish between the different types of data used for calculating RISE and the different levels of replications taking place. As explained in Chapter 2, RISE does not follow a known sampling distribution and thus a statistical significance test of RISE cannot be easily conducted. Instead, an empirical sampling distribution of RISE may be created using a bootstrapping procedure. Five hundred replications of item responses to each item, all of which fit the model, are generated to create the empirical sampling distribution of RISE. This empirical sampling distribution can then be used for the purposes of statistical significance testing. For consistency, even though items 44-50 were generated to fit the model in the bootstrapping procedure, these items were not included in the calibration of θ. In an applied setting, the bootstrapping procedure would involve simulated data and would be used to assess the statistical significance of RISE for the “real” data.

However, in this study, the “real” data are also simulated. Thus, there are two different types of simulations being performed. For the remainder of the study, the “real” data, which is only created once per replication of the study, will be called the “simulated” data. The replication to create this simulated data will be called a “macro-
replication.” The simulated data used to create the empirical sampling distribution of RISE will be called the “bootstrap” data. Each of the 500 replications used to create the empirical sampling distribution will be called “micro-replications.” Thus, within every 1 macro-replication of this study, there will be 500 micro-replications.

Within each macro-replication, RISE was calculated for all 50 simulated items on the test, using equation 2.18. The parametric proportion correct \( \hat{P}_j \) at each evaluation point \( Q \) was calculated simply by subtracting the estimated difficulty of the item from the \( \theta \) value at each \( Q \) and converting the logits into probabilities. The non-parametric proportion correct \( \hat{P}_{\text{non}, ji} \) was obtained by smoothing the observed proportion correct at each observed total score.

**Calculating RISE: Smoothing the simulated data.** Smoothing was conducted using one of two smoothing methods, depending on the condition. The first method was Hanning using Equation 2.13, with examinees grouped by their estimated \( \theta \) prior to smoothing. For simplicity, this method will be called “Hanning by \( \theta \).” Given the one-to-one correspondence of estimated \( \theta \)’s and raw total scores, grouping examinees by \( \theta \) is equivalent to grouping examinees by total score. Given this grouping prior to smoothing, there were as many data points to be smoothed as there were estimated \( \theta \)’s. This means that \( Q \) in Equation 2.18 was the number of unique estimated \( \theta \)’s. Given that examinees were grouped by their estimated \( \theta \), the data points to be smoothed were the observed proportion correct at each estimated \( \theta \). Depending on the level of the smoothing amount factor, data were run through the smoother multiple times. Once the proportion correct at each \( Q \) was smoothed the specified number of times, RISE was calculated for each item. As specified in Equation 2.18 with the term \( f(\theta) \), the squared difference between the
parametric and non-parametric proportion correct was weighted by the density at each evaluation point $Q$. Thus, regions of the estimated $\theta$ distribution containing more examinees were weighted more heavily when calculating RISE.

From a practical point of view, it makes sense to weight proportion correct by the density of $\theta$ at each evaluation point $Q$. For example, a practitioner may find that an item contains significant graphical misfit in one section of the $\theta$ scale. However, if there are no or few people with $\theta$’s on that part of the scale, then from a practical standpoint, the misfit is not consequential for that sample of examinees. It would be imprudent to throw that item out, given that there are no examinees for which it is functioning inappropriately.

The second smoothing method was kernel smoothing with a Gaussian function. Examinees were first grouped by their estimated $\theta$. Thus, as in the Hanning by $\theta$ method, the number of data points to be smoothed equaled the number of unique estimated $\theta$’s. Then, using Equations 2.15 and 2.16, smoothing weights were calculated for every data point and were used to smooth the observed proportion correct at each estimated $\theta$. Then, RISE was calculated using Equation 2.18. As with the Hanning by $\theta$ method, differences between the parametric and non-parametric curve were weighted by the density of $\theta$. As with both prior smoothing methods, the amount of smoothing used for Kernel smoothing depended on the level of the smoothing amount factor.

In both the Hanning by $\theta$ and Kernel smoothing with a Gaussian function, examinees were grouped by their estimated $\theta$, not some transformation of the rest score or total score, as in previous studies. Thus, it may seem as if it is not a truly non-parametric technique, given that examinees were grouped based on a parameter estimate.
However, given that the model of choice for this dissertation is the Rasch model, grouping by estimated θ is equivalent to non-parametric grouping techniques. As specified in Chapter 2, there is a one to one equivalence between raw total scores and estimated θ’s in the Rasch model. All examinees with the same total score will have the same estimated θ. Thus, grouping examinees by their raw total scores is equivalent to grouping them by their estimated θ’s. In previous studies (Douglas & Cohen, 2001; Wells & Bolt, 2008), the 2PL was the model of choice. With the 2PL model, total scores are not sufficient statistics for estimating parameters. Thus, grouping by rest score, total score, or some transformation of them, will not be equivalent to grouping by estimated θ’s.

**Calculating RISE: Smoothing amount.** Within each smoothing method, nine amounts of smoothing were used. The amounts were chosen to represent a wide range of possible smoothing amounts, from essentially no smoothing at all to a great deal of smoothing. As mentioned previously, the Hanning procedure can be used in multiple iterations to yield a smoother and smoother curve, with the number of iterations dictating the smoothness of the curve. For the Hanning by θ method, 0, 5, 10, 15, 20, 25, 30, 35, and 40 iterations were used. The 0 means that the data were not smoothed at all using the Hanning procedure. Instead, examinees were only grouped by their estimated θ’s. If RISE functions as well with no smoothing as it does with smoothing, then using RISE as a measure of item fit would be simplified in that smoothing would not even be necessary.

For kernel smoothing, there were no iterations of smoothing used to determine the level of smoothness. All the smoothing was conducted in a single step. Instead, the level of smoothness was dictated by the bandwidth parameter $h$, as seen in Equation 2.15. A smaller bandwidth parameter yields more rapidly decreasing smoothing weights and thus
a less smoothed curve. Bandwidth parameters of 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8 were chosen.

For both smoothing methods, it was hypothesized that there would be a curvilinear relation between smoothing amount and the power of RISE. At the lowest smoothing amounts, power would be relatively low due to under smoothing. As the amount of smoothing increased, power would increase. At a certain smoothing threshold, power would actually begin to decrease due to over smoothing. Thus, there would be an ideal amount of smoothing, with smoothing amounts smaller or larger than the ideal yielding lower amounts of power.

**Calculating RISE: Bootstrapping.** As explained in Chapter 2, RISE does not follow a known sampling distribution. Thus, it is not enough to simply calculate RISE for each simulated item, as there is no significance test for it. Instead, a bootstrapping procedure was conducted to create an empirical sampling distribution for RISE. This procedure was conducted as described in Chapter 2, with 500 micro-replications to create the empirical sampling distribution of RISE for each item. The amount of smoothing within each micro-replication matched the amount of smoothing within the macro-replication.

Once the empirical sampling distribution was created, an alpha level of .05 was chosen, so the 95th percentile of the RISE sampling distribution for each item was chosen as the critical value. If an item’s RISE value exceeded the 95th percentile critical value obtained from the bootstrapping procedure, then that item was flagged as statistically significantly misfitting for that macro-replication. This was done for all 50 items on the test.
Calculating Power and Type 1 Errors: Infit/Outfit and RISE

There were 500 macro-replications conducted for each condition. Given that there were 3 levels of sample size, 2 levels of smoothing method, and 9 levels of smoothing amount, this yielded 54 unique conditions. Criteria for comparison across the conditions were the type 1 error rates and power of standardized Infit, standardized Outfit, and RISE.

A very important property of a fit statistic, or any statistic for that matter, is that it maintains the nominal type 1 error rate. If a fit statistic has inflated type 1 error rates, then a practitioner will deem good items as misfitting more often than he/she should, which has economic and practical implications. If a fit statistic has deflated type 1 error rates, then that will negatively impact the power of that statistic, which will be discussed subsequently. Given that the first 43 items on the test were generated to fit the Rasch model, these items were used to assess type 1 error rates of the fit statistics across conditions. These were calculated for each of the 43 items by simply dividing the number of macro-replications in which the item was flagged as misfitting by the total number of macro-replications, which was 500. This represents the proportion of macro-replications in which a good item was incorrectly identified as misfitting.

The power of a fit statistic is also consequential. A good fit statistic will have enough power to adequately detect measurement disturbances in the data. If a fit statistic has poor power, then a practitioner will not be able to identify poor items. In addition, a fit statistic should be sensitive to many different types of measurement disturbances. Given that items 44 through 50 were simulated to convey various measurement disturbances, they were used to assess the power of the three fit statistics. Power was
calculated for each item by dividing the number of macro-replications in which the item was flagged as misfitting by the total number of macro-replications. This represents the proportion of macro-replications in which a bad item was correctly identified as misfitting the Rasch model.

RISE will be compared to Infit and Outfit in terms of both type 1 error rates and power.

**Justification for the Simulation Design**

The number of items (50) was chosen to represent a medium length test, a length which is common in educational testing. Item difficulties were chosen to cover a broad range of values. These values were chosen primarily to examine the impact of item difficulty on the type 1 error rates of Infit and Outfit. As discussed in chapter 2, previous literature has shown that the alignment of an item’s difficulty with the examinee population can drastically affect type 1 error rates of Outfit, and especially Infit. A wide range of item difficulties was chosen to confirm previous findings.

Seven different types of item misfit were examined in this study. A major purpose of this study was to compare the performances of RISE to Infit and Outfit across different types of item misfit. As discussed in chapter 2, Infit and Outfit are measures of the degree to which an item’s discrimination is different than the average discrimination of the other items. RISE is a quantification of the difference between the parametric and non-parametric IRF. Given that these fit statistics are calculated and interpreted differently, they may be sensitive to differing types of misfit. Thus, 7 types of misfit were simulated in order to explore this issue in depth.
The 43 fitting items and 7 misfitting items were placed on the same test. This was done so that type 1 error rates and power could be examined simultaneously, as opposed to being examined in separate simulation studies. This significantly reduced computing time. Of course, the presence of misfitting items could affect the type 1 error rates for the fitting items. However, given that items 44-50 were not included in the estimation of \( \theta \), their misfit did not contaminate the fitting items.

Three levels of sample size were chosen to represent small, medium, or large samples for Rasch modeling. Sample size should not affect type 1 error rates. However, it will affect power, as is the case with any statistical significance test. Given that power was one of the two criteria for comparison in this study, it was logical to vary sample size. For example, perhaps RISE performs better than Infit and Outfit at low sample sizes. Yet, when sample sizes are large, power is so large across all three fit statistics that it is inconsequential as to which one is used.

Two types of smoothing were investigated in this study. While there is literature describing smoothing techniques for item response data, there is little evidence to suggest that one smoothing method is more effective than another. Given that some smoothing techniques are much simpler, both conceptually and computationally, than others, it would be advantageous to find that the simpler smoothing techniques are just as effective as the more difficult ones. Thus, two smoothing methods were chosen in order to investigate whether the type of smoothing has an effect on the type 1 error rates and power of RISE. Given that computation of Infit and Outfit does not involve smoothing, their values will remain the same regardless of how much smoothing is conducted for the computation of RISE.
Within each smoothing method, nine amounts of smoothing were chosen. Conventional wisdom is that under-smoothing will lead to too much chance roughness in the data, whereas over-smoothing will obscure meaningful patterns in the data. Instead, the goal should be to find a “happy medium” of smoothing. However, there is no research to suggest what that happy medium should be, or even if the notion of finding a happy medium is the correct approach. Thus, nine differing amounts of smoothing were chosen to investigate this issue. If the notion of a happy medium is correct, then power should initially increase as the amount of smoothing increases, reach a plateau, and eventually decrease for the largest smoothing amounts.
CHAPTER 4

Results

Research Question 1

Research Question 1: Type 1 errors. Research Question 1 pertained to the effect of sample size on the type 1 error rates and power of RISE, Infit, and Outfit. Ideally, for any statistical significance test, type 1 error rates should not be affected by sample size. Power, of course, should increase as sample size increases. Table 1 contains type 1 error rates for the Hanning RISE across all nine smoothing amounts; type 1 error rates were averaged across the 43 items that were generated to fit the Rasch model.

As can be seen in Table 1, type 1 error rates for Hanning RISE are near the specified value of .05 across all smoothing amounts and sample sizes. The small departures from .05 are likely due to sampling error. If more macro and micro replications had been used in the study, the values would be even closer to .05 than they were in this study. Table 2 contains type 1 error rates for Kernel RISE.

As with Hanning RISE, type 1 error rates for Kernel RISE were consistently near .05 across all sample sizes and smoothing amounts. However, the type 1 error rate for the condition when sample size was 500 and the bandwidth was 0.5 was not reported in Table 2. This is because of a technical error with this condition that rendered the results for this condition inaccurate. The condition was re-run multiple times and the results were still wildly inaccurate. There was no apparent reason for the error associated with this condition. Therefore, the type 1 error rate for this condition is not reported. As will be seen, power for this condition was also not reported. Thankfully, the loss of this single
condition did not in any way obscure the patterns of results, nor did it affect the answering of the four research questions of this study.

The type 1 error rates for RISE were expected. Recall that RISE does not follow a known sampling distribution. Thus, bootstrapping was used to create an empirical null sampling distribution of RISE and cut-off values were determined by identifying the 95th percentile of RISE values for each item across micro-replications. Because of this, the type 1 error rate must be .05. Any systematic departures from .05 would indicate some type of programming or software error, not an issue with the RISE statistic, as was the case with the condition described in the previous paragraph.

Table 3 contains type 1 error rates for Infit and Outfit. Infit and Outfit do not have type 1 error rates that are consistent with the nominal level of .05. Instead, they are deflated, especially for Infit. This was the case for all three sample sizes. This also was an expected result. As discussed in Chapter 2, several studies have shown that item targeting has a significant impact on type 1 error rates for Infit and Outfit. Type 1 error rates for Infit and Outfit will be near .05 for items with difficulties near the ability of the examinees. Type 1 error rates will rapidly become deflated as items become too difficult or too easy for the examinees. This was observed in this study. Items with difficulties near 0 were associated with Infit and Outfit type 1 error rates of .05, because the examinees had a mean ability of 0. For the easier and harder items, type 1 error rates were deflated. For the easiest and hardest items on the test, type 1 error rates for Infit and Outfit were near 0. This is why the type 1 error rates for Infit and Outfit, averaged across items, were below the specified value of .05. This same phenomena did not occur with RISE. There was no relation between item difficulty and type 1 error rates for RISE.
**Research Question 1: Power.** Given that each of the seven misfitting items were generated according to a different model, it did not make sense to average power rates across the seven items. Tables 4 and 5 contain power results for Infit and Outfit, respectively. Differences in power between the items will be discussed in a subsequent section. As expected, power of both Infit and Outfit almost universally increased as sample size increased. There were two exceptions. The first was Infit for item 45. In comparing sample sizes of 200 and 500, power actually slightly decreased. The other exception was Outfit with item 49. Outfit for item 49 did not increase when the sample size increased from 500 to 1,000. However, this was because power was 1 in both conditions. In fact, for both Infit and Outfit on items 46, 47, and 49, power barely changed when the sample size increased from 500 to 1,000. This is because power is, of course, bounded at 1. Once perfect power is achieved, any increase in sample size cannot yield higher rates of power.

Tables 6-12 contain power results for Hanning RISE. Each table contains results for one of the seven misfitting items. As stated in the previous paragraph, differences in power between items and between smoothing amounts will be discussed in subsequent sections. Similar to Infit and Outfit, the power of Hanning RISE almost always increased as sample size increased, regardless of the item or the amount of smoothing. Also similar to Infit and Outfit, some items exhibited only a very small increase in power as sample size increased from 500 to 1,000. This was also due to the upper bound of 1 for power. Tables 13-19 contain power results for Kernel RISE. The pattern of results was consistent with the pattern observed for Infit, Outfit, and Hanning RISE. Increases in sample size
almost always yielded an increase in power, regardless of the item or the amount of smoothing.

**Research Question 2**

Research Question 2 pertained to the effect of smoothing amount on the type 1 error rates and power of RISE. As was discussed in the previous section, type 1 error rates were consistently near .05, regardless of the amount of smoothing. Again, this was expected because the null sampling distribution of RISE was created empirically. Thus, the focus of this section will be on the effect of smoothing on the power of RISE.

**Research Question 2: Hanning.** Because each of the seven misfitting items was generated using a different model, power will be discussed for each misfitting item individually. Figure 8 displays power results for item 44, which was called the “wavy” item. As can be seen, even for the largest sample size, power was quite poor. Power was highest for the lowest smoothing amount, which was no smoothing at all in the case of Hanning RISE. As the amount of smoothing increased from 0, 5, and 10 smoothing iterations, power decreased, rapidly in the 500 and 1,000 sample size conditions. From 15 to 40 smoothing iterations (the largest amount of smoothing), power essentially plateaued. Additional smoothing iterations past 15 had little effect on the power of RISE. Thus, a “happy medium” amount of smoothing did not exist for this item; smoothing actually had a detrimental effect on power, regardless of the amount.

Figure 9 displays power results for item 45, which was called the “big dip” item. As can be seen, the power of RISE was significantly affected by the smoothing amount in the sample size of 200 condition. Power increased drastically from 0 to 5 smoothing iterations (from .558 to .856). From 5 to 40 smoothing iterations, power decreased in a
fairly linear fashion (from .856 to .414). Thus, power ranged from quite good to poor when sample size was 200, depending on the amount of smoothing. The same trend can be observed in the sample size of 500 condition. The trend is much less drastic however, given that power was so high across all smoothing amounts. When sample size was 1,000, power was so high that the smoothing amount was inconsequential.

Figure 10 displays power results for item 46, which was generated according to a 3PL model. Across all 3 sample sizes, power was lowest when there were 0 smoothing iterations. Moving from 0 to 5 smoothing iterations was associated with the largest increase in power, regardless of sample size. Power peaked at 20 smoothing iterations when sample size was 200 or 500 and at 10 smoothing iterations when sample size was 1,000. Across sample sizes, after power peaked, it more or less plateaued, with increases in smoothing only marginally affecting power. As with item 45, this pattern of power is less noticeable when sample size was 1,000, given that power was so high.

Figure 11 displays power for item 47, which was generated using a 2PL model. When sample size was 200, power was very poor across all smoothing amounts. However, there was a pattern. Power increased markedly from 0 to 5 smoothing iterations, increased slightly from 5 to 10, and decreased with every subsequent increase in smoothing. When sample size was 500 or 1,000, power followed a similar pattern as when sample size was 200, although power peaked at 5 smoothing iterations instead of 10. The pattern of power is more easily recognizable when sample size was 500 or 1,000, given that power was so poor across all smoothing amounts for the sample size of 200. Unlike items 44 and 46, there was no plateauing effect associated with larger amounts of smoothing. Instead, power reached a peak and then decreased with additional smoothing.
For example, when the sample size was 1,000, power peaked at .922 for 5 smoothing iterations and was .178 for 40 smoothing iterations.

Figure 12 displays power for item 48, which was generated using a 4PL model. Across sample sizes, the patterns of power were very similar to the pattern observed for item 47. Power started low when there was no smoothing, increased drastically up to a certain smoothing amount (5 smoothing iterations in this case), and then consistently decreased as the amount of smoothing increased.

Figure 13 displays power results for item 49, which was generated using a hyperbolic cosine model. Opposite of the results for item 44, power was generally quite good regardless of the sample size or the amount of smoothing. Unlike previous items, the amount of smoothing did not have much of an effect on power. When sample size was 200 or 500, power was lowest when there was no smoothing. When sample size was 200, power increased up until 15 smoothing iterations, then very marginally decreased with further smoothing. When sample size was 500, power was essentially unaffected by smoothing amount, other than increasing noticeably from 0 to 5 smoothing iterations. When sample size was 1,000, power was essentially perfect across all smoothing amounts.

Lastly, Figure 14 displays power results for item 50, which was called the “flat middle” item. When sample size was 200 or 500, power was lowest when there was no smoothing. Power peaked at 5 smoothing iterations when sample size was 200 and 10 smoothing iterations when sample size was 500. In both these sample size conditions, power then decreased as power increased, but more substantially when sample size was
200. As with items 45 and 49, when sample size was 1,000, power was perfect or nearly perfect across all smoothing amounts.

Across the seven misfitting items, three general patterns emerged in regards to the relation between smoothing amount and power. The first pattern was the “happy medium” pattern. With this pattern, power increases as smoothing increases, reaches a peak, and then begins to decrease as smoothing continues to increase. Items 45, 47, 48, and 50 followed this pattern, although the pattern was obscured for items 45 and 50 when the sample size was 1,000, given that power so high across all the smoothing amounts. For these four items, both undersmoothing and oversmoothing were real threats to power. The optimal amount of smoothing for these items was either 5 or 10 smoothing iterations, depending on the item and the sample size. The second pattern was the plateau pattern. This pattern was characterized by an initial increase in power as smoothing increased and then a tapering off of power as smoothing continued to increase. Items 46 and 49 followed this pattern. With this pattern, undersmoothing was a threat to power, but oversmoothing was not. Smoothing had an effect on power for these items up to 10 smoothing iterations. Lastly, the third pattern was the “no smoothing” pattern. Only Item 44 followed this pattern. For the case of item 44, any amount of smoothing actually decreased power. Therefore, for all but item 44, the optimal amount of smoothing iterations was between 5 and 10 iterations.

Research Question 2: Kernel. Figures 15-21 display power results for Kernel RISE for each of the seven misfitting items. While the actual power values may differ across Hanning RISE and Kernel RISE, the patterns of power and smoothing were often similar across the two smoothing techniques. As can be seen, items 45, 47, and 48 all
followed the “happy medium” pattern for Kernel RISE, just as they did for Hanning RISE. For item 45, when sample size was 200, power peaked at a bandwidth of 0.2. For item 45, when sample size was 500 or 1,000, power was perfect or nearly perfect up until a bandwidth of 0.5. For items 47 and 48, across all sample sizes, power peaked at a bandwidth of 0.2.

As with Hanning RISE, items 46 and 49 both followed the “plateau” pattern for Kernel RISE. For both items, power increased most noticeably on the low end of the bandwidth scale. When moving to the upper end of the bandwidth scale, power barely changed. Again, when the sample size was 1,000, this pattern essentially did not exist for item 49, given that power was perfect or nearly perfect across all bandwidth parameters.

There were some differences in the patterns of results between Hanning and Kernel RISE. For Kernel RISE, item 44 did not follow the “no smoothing” pattern. Recall that smoothing immediately worsened power for Hanning RISE. For Kernel RISE, this was not the case. Instead, power peaked at a bandwidth parameter of 0.1, then decreased for subsequent bandwidth parameters. Thus, for Kernel RISE, item 44 also followed the “happy medium” pattern.

In a similar vein, Kernel RISE for item 50 did not particularly follow the “happy medium” pattern as Hanning RISE did. As can be seen in Figure 21, when sample size was 200, the power of Kernel RISE increased up until a bandwidth of 0.3. It then plateaued from bandwidths of 0.3 to 0.6, and then eventually decreased for bandwidths of 0.7 and 0.8, but only slightly. When sample size was 500, power peaked at a bandwidth of 0.2 and then plateaued. When sample size was 1,000, power was perfect or nearly
perfect across all bandwidths. Thus, for Kernel RISE, item 50 tended to more closely
follow the “plateau” pattern instead of the “happy medium” pattern.

Finally, both Hanning and Kernel RISE followed the “happy medium” pattern for
item 45, as stated earlier. However, Figures 9 and 16 look quite different when sample
size was 500 or 1,000. Larger amounts of smoothing had a much greater impact on power
for Kernel RISE than Hanning RISE. Even when sample size was 1,000, the power of
Kernel RISE became quite low for the largest bandwidth parameters. For Hanning RISE,
power was essentially perfect across all smoothing amounts when the sample size was
1,000. When the sample size was 500, the power of Hanning RISE did decrease for the
larger smoothing amounts, but much less drastically than Kernel RISE. Thus,
oversmoothing was a much bigger threat to power for Kernel RISE than Hanning RISE
for item 45.

Research Questions 3 and 4

Research Question 3 pertains to the effect of smoothing technique on the type 1
error rates and power of RISE. Part of this research question was already addressed in the
previous sections regarding Research Question 2. A was shown in tables 1 and 2, type 1
error rates for both Hanning and Kernel RISE were near the nominal alpha level of .05,
regardless of sample size or smoothing amount. Thus, smoothing technique did not have
an effect on type 1 error rates. As was discussed in the subsequent section, the patterns of
smoothing amount and power for Hanning RISE and Kernel RISE were similar for some
items and different for others. However, comparisons of Hanning and Kernel RISE in
terms of peak performance across smoothing amounts have not yet been addressed.
Those comparisons will be made while also addressing Research Question 4.
Research Question 4 pertained to whether RISE has different rates of power than Infit and Outfit across various types of misfitting items. This question may be answered by examining Figures 22 through 28. Each figure pertains to one of the misfitting items. On each figure, sample size is on the x-axis and power is on the y-axis. Infit and Outfit are represented by the triangle and square symbol, respectively. As can be seen, Figures 22 through 28 do not display power results for every smoothing amount of Hanning and Kernel RISE. Instead, they show the peak power across all smoothing amounts for Hanning and Kernel RISE. These are represented by the circle and star symbols, respectively. For the exact power results, consult Table 20. Bolded values indicate the highest power in that condition across the four fit statistics.

Figure 22 depicts the comparison of the four fit statistics for item 44, the wavy item. As can be seen, power was quite poor, regardless of sample size or the fit statistic being used. Across all three sample sizes, Kernel RISE outperformed Hanning RISE. When sample size was 200, Kernel RISE performed the best of the four statistics, Hanning RISE and Outfit were almost identical, and Infit performed the worst. When sample size was 500 or 1,000, Outfit actually performed the best. This was a surprising result. Recall that item 44 is the wavy item depicted in Figure 1. The true generating model of the wavy item produces a discrimination equal to 1. Recall that Infit and Outfit are a measure of whether the discrimination of an item is different than the average discrimination of the remaining items. RISE, instead, is a literal quantification of the difference between the observed IRF and the theoretical IRF. Given these two facts, it was hypothesized that Infit and Outfit would perform worse than RISE. Instead, for two of the three sample sizes, the reverse was true. Power of Infit and Outfit was larger than
was expected and the power of Hanning and Kernel RISE was much lower than expected. However, as can be seen, power was very similar across all sample sizes for the four statistics.

Figure 23 depicts the comparison of the fit statistics for item 45, the big dip item. When sample size was 200, Hanning RISE performed the best. Kernel RISE performed slightly worse. Outfit was noticeably worse than either RISE statistics, and Infit performed terribly. When the sample size was 500 or 1,000, Hanning RISE, Kernel RISE, and Outfit all had either perfect or nearly perfect power. Infit had abysmal power, even when the sample size was 1,000.

Figure 24 depicts the comparison of the fit statistics for item 46, the 3PL item. When the sample size was 200, Kernel RISE noticeably outperformed Hanning RISE. However, when the sample size was 200 and 500, Infit and Outfit outperformed the two RISE statistics. Differences between Infit/Outfit and the RISE statistics were largest when the sample size was 200. As sample size increased, the differences became smaller. When sample size was 1,000, all four statistics had either perfect or nearly perfect power.

Figure 25 depicts the comparison of the fit statistics for item 47, the 2PL item. Across sample size, Hanning RISE performed much better than Kernel RISE. When sample size was 200 Infit clearly performed the best of the four statistics. When sample size was 500 or 1,000, Infit and Outfit both had perfect or nearly perfect power. Power did not reach acceptable levels until a sample size of 1,000 for Hanning RISE and never reached an acceptable level for Kernel RISE.

Figure 26 depicts the comparison of the fit statistics for item 48, the 4PL item. Kernel RISE performed better than Hanning RISE when the sample size was 200, but the
reverse was true when the sample size was 500 or 1,000. Regardless of sample size, Outfit performed the best, much better than the other statistics when the sample was 200 or 500. Like with item 45, Infit performed very poorly, even with the largest sample size.

Figure 27 depicts the comparison of fit statistics for item 49, the hyperbolic cosine item. As can be seen, power was quite good, even when the sample size was small. Kernel RISE outperformed Hanning RISE for this item when the sample size was 200. Further, when the sample size was 200, Outfit performed noticeably better than the other statistics. When the sample size was 500 or 1,000, all the fit statistics had either perfect or nearly perfect power.

Lastly, Figure 28 depicts the comparison of fit statistics for item 50, the flat middle item. Like the previous item, Kernel RISE outperformed Hanning RISE when the sample size was 200. Infit performed the best when the sample size was 200 and Outfit was clearly the worst. When the sample size was 500, Infit, Hanning RISE, and Kernel RISE performed similarly well while Outfit was still clearly worse. When the sample size was 1,000, all four statistics had perfect or nearly perfect power. Unlike many of the previous items where Outfit was the top performer, Outfit had the lowest power across sample sizes for item 50.

In examining power for every item by sample size, the four fit statistics were compared 21 times. Outfit was the top performing fit statistic across these comparisons. Outfit had the highest power, or was tied for the highest power, in 13 of the 21 comparisons. Infit performed second best, in terms of peak performance, as it had the highest or tied for highest power in 10 of the comparisons. However, for two items (45 and 48), it did have drastically worse power than the other three statistics. Hanning and
Kernel RISE performed equally the worst, with only 5 comparisons where they had either the highest power or tied for the highest power. There was not a single item where either RISE statistic was consistently the best performer. While Hanning RISE and Kernel RISE yielded different rates of power, the results were very similar on all items but item 47.
CHAPTER 5

Discussion

As a reminder to the reader, the four research questions that this study addressed were as follows:

Research Question 1: How does sample size affect the type 1 error rates and power of RISE and standardized versions of Infit and Outfit?

Research Question 2: How does the amount of smoothing affect type 1 error rates and power of RISE?

Research Question 3: Does the type of smoothing technique affect type 1 error rates and power of RISE for the Rasch model?

Research Question 4: Does RISE have different power rates than standardized Infit and Outfit across various types of item misfit?

A simulation study was conducted to answer these questions. Scored item responses were generated for a 50-item test. The first 43 items were generated using the Rasch model, in order to test type 1 error rates of the fit statistics. Items 44-50 were each generated using a different model that would cause a lack of fit to the Rasch model, in order to test the power of the fit statistics. Sample size was the first factor and was manipulated to have three levels: 200, 500, or 1,000 examinees. Smoothing technique, when calculating RISE, was the second factor and had two levels: Hanning and Kernel smoothing with a Gaussian function. With both techniques, smoothing was done on the proportion correct at each ability level, not on the raw item responses. The final factor was smoothing amount and had nine levels, ranging from very low to very high amounts of smoothing. For Hanning smoothing, the amount of smoothing was controlled by the
number of smoothing iterations used. For Kernel smoothing, the amount of smoothing was controlled by the bandwidth parameter, with larger bandwidths yielding high levels of smoothing. Smoothing amount was nested within smoothing technique, which was crossed with sample size, yielding 54 unique conditions. Each condition was replicated 500 times. In this discussion below, each research question will be addressed in order.

**Research Question 1**

A stable null distribution is a desirable quality of any statistic. Without it, statistical significance tests of that statistic are either complicated or impossible. The stability of the null distributions of Infit, Outfit, Hanning RISE, and Kernel RISE across sample sizes was investigated by assessing type 1 error rates. As was shown in Tables 1, 2, and 3, type 1 error rates for all four statistics were invariant across sample size. For both RISE statistics, type 1 error rates were always near the nominal level of .05. As discussed previously, this was expected, as the null distribution was empirically created for RISE. For Outfit and especially Infit, type 1 error rates were deflated. As prior research has shown, this was due to presence of items with difficulties that were far from the mean ability of the examinees. Many studies (DeMars, 2017; Karabatsos, 2000; Wang & Chen, 2005; Wu & Adams, 2013) have shown that type 1 error rates for Infit and Outfit will be near .05 for well targeted items and will become decreasingly less than .05 as items become too easy or too hard. This study confirmed this property of Infit and Outfit and the results were unsurprising.

As for the power of the fit statistics, it was expected that power would increase as sample size increased, as is the case for virtually all statistics. This was confirmed in this study. Almost universally, power increased for all the fit statistics as sample size
increased, regardless of the item. The only cases where an increase in sample size did not yield an increase in power was when power was already perfect for the lower sample size. The results were again not surprising. The more interesting aspects of sample size in this study was its effects on comparisons of power across smoothing amounts and on comparisons of the four fit statistics. These will be addressed in the subsequent sections of the discussion.

**Research Question 2**

As was shown in Chapter 4, smoothing amount did not have an effect on type 1 error rates for either Hanning RISE or Kernel RISE. This was an expected result. Smoothing amount did affect the power of Hanning RISE and Kernel RISE. The relationship between smoothing amount and power for each item tended to follow one of three patterns: “no smoothing”, “happy medium”, and “plateau”.

The “no smoothing” pattern, which was only observed with Hanning RISE on item 44, was characterized by any amount of smoothing leading to a decrease in power. The “happy medium” pattern was observed with Hanning RISE for items 45, 47, 48, and 50 and with Kernel RISE for items 44, 45, 47, and 48. This pattern was characterized by an initial increase in power as smoothing increased, a peak in power, and then a decrease in power as smoothing continued to increase. The peak in power tended to happen at relatively low amounts of smoothing, usually 5-10 smoothing iterations for Hanning RISE and a bandwidth of 0.2 for Kernel RISE. The “plateau” pattern was observed with Hanning RISE for items 46 and 49 and with Kernel RISE for items 46, 49, and 50. This pattern was characterized by an initial increase in power as smoothing increased, and then
an eventually tapering off such that further smoothing did not have a meaningful effect on power.

Thus, the amount of smoothing did not have a universal effect on power. Instead, the relationship between smoothing amount and power depended on the type of item misfit. Additionally, sample size sometimes moderated this relationship. For examples, on items 46, 49, and 50, when sample size was 1,000, both Hanning and Kernel RISE had perfect or nearly perfect power across all smoothing amounts. This was also the case on item 45 for Hanning RISE. In these instances, the amount of smoothing actually was inconsequential. These items would be flagged as misfitting regardless of how much smoothing was conducted. Therefore, the relationship between smoothing amount and power depended not only on the type of misfit, but also on the sample size.

Figures 29 to 35 were created to help determine why each item followed one of the three power patterns. There is a figure for each misfitting item. Each figure depicts the observed IRFs from one replication using three different smoothing amounts, which constitutes relatively low, medium, and high amounts of smoothing. Given that the patterns of results were generally similar across the two smoothing techniques, and to reduce the number of graphs, Figures 29 to 35 are only for Kernel smoothing.

Figure 29 shows the smoothed IRFs for item 44, the “wavy” item. As can be seen, as the bandwidth parameter increases, the IRF becomes increasingly smooth. When the bandwidth was 0.05, which constituted almost no smoothing at all, the IRF was excessively jagged. When the bandwidth was 0.2, the jaggedness was largely removed from the IRF, but it still retained the systematic waviness. When the bandwidth was 0.8, all of the systematic waviness inherent to the true generating IRF was removed from the
observed IRF. Instead, the IRF when the bandwidth was 0.8 looks quite similar to a Rasch implied IRF, which explains why power was lowest for high bandwidth parameters on item 44.

Figure 30 shows the smoothed IRFs for item 45, the “big dip” item. As with item 44, the observed IRF using a bandwidth of 0.05 was quite jagged. When a bandwidth of 0.2 was used, most of the jaggedness was removed, but the systematic dip in the middle of the ability continuum was preserved. When the bandwidth was 0.8, the dip was still present, but was manifested to a lesser degree than the true underlying IRF would imply. Thus, it again makes sense as to why power was lowest for the large bandwidth parameters.

However, one might expect that power for this item would be highest when the bandwidth was 0.05, given the extreme jaggedness of the IRF. From Figures 30, it does appear that the observed IRF with a bandwidth of 0.05 should fit worse than the observed IRF with a bandwidth of 0.2. In fact, the actual value of RISE was larger when the bandwidth was 0.05 then it was when the bandwidth was 0.2 for item 45. However, the critical value of RISE was also much larger when the bandwidth was 0.05. Thus, power was significantly higher when the bandwidth was 0.2, even though the actual RISE value was smaller than when the bandwidth was 0.05.

This can be explained by recalling the process for conducting a statistical significance test for RISE. Use item 45 as an example. Within each macro-replication of the study, item 45 was generated to appear as it does in Figure 30. Smoothing was performed to create item 45’s observed IRF (as shown in Figure 30), and RISE was calculated by quantifying the difference between the item 45’s observed IRF and item
45’s theoretical IRF (provided by Winsteps) using Equation 2.18. But it cannot immediately be known whether the RISE value calculated for item 45 in this macro-replication was statistically significant. As laid out in the method, the null sampling distribution for RISE for item 45 had to first be created.

This was done by generating item responses to item 45 using item 45’s estimated difficulty. These item responses were generated to fit the Rasch model. These item responses were then smoothed to create a null observed IRF for item 45. Winsteps was used to attain the theoretical IRF for these responses to item 45. A null RISE value was then calculated using the null observed IRF and the theoretical IRF. This process was repeated 499 more times (which were called micro-replications in chapter 2) to create a null sampling distribution of RISE for item 45. It was a null sampling distribution of RISE because all 500 RISE values were calculated as if item 45 fit the Rasch model (even though it actually did not). The RISE value associated with the 95th percentile of this null distribution for item 45 became the critical value for item 45. This critical value was then used to determine if item 45 statistically significantly misfit the Rasch model, within each macro-replication of the study.

The key point in the process, as it pertains to power, was the smoothing of the observed IRF for item 45 within each micro-replication. When a small bandwidth parameter was chosen, like 0.05, then the observed IRF for item 45 in each micro-replication looked very jagged. Even though the observed IRF for item 45 within a micro-replication was generated according to the Rasch model, the jaggedness caused a fairly large discrepancy between the observed IRF for item 45 and the theoretical IRF for item 45 provided by Winsteps. This led to a relatively large RISE value within each
micro-replication. When the 500 RISE values across micro-replications for item 45 were formed into a null sampling distribution, they constituted a null sampling distribution of RISE values created from very jagged observed IRF’s. This meant that the 95th percentile of this null distribution was large, and thus the critical value was large. This meant that the actual RISE value for item 45 within a given macro-replication also had to be large for it to be flagged as statistically significant. RISE was essentially being penalized for using such a little amount of smoothing when creating the empirical sampling distribution.

When a larger bandwidth was used instead, like 0.2, the observed IRF for item 45 within each micro-replication was smoother and better resembled the theoretical IRF. This led to smaller RISE values for item 45 within each micro-replication. In turn, this led to a smaller critical value for item 45 than when the bandwidth was 0.05. This meant that a smaller value of RISE could be flagged as statistically significant when using a bandwidth of 0.2 than when using a bandwidth of 0.05. This smaller critical value for item 45 using a bandwidth of 0.2 was much more influential than the fact that observed IRF when using a bandwidth of 0.05 appeared to fit worse than the observed IRF when using a bandwidth of 0.2, as depicted in Figure 30. Thus, power for Kernel RISE was not highest when the smoothing amount was smallest, despite what appearances in Figures 28 and 29 might suggest. This phenomenon was true of every item when using Kernel smoothing and all items except for item 44 when using Hanning smoothing.

Figure 31 shows smoothed IRFs for item 46, with was the 3PL item. Once again, the IRF was quite jagged when using a bandwidth of 0.05. When the bandwidth was 0.2, much of the jaggedness was removed, but the IRF was still wavier than would be
expected for a 3PL item. The IRF using a bandwidth of 0.8, as can be seen, was noticeably flattened in comparison to the IRFs when the bandwidth was 0.05 or 0.2. The observed IRF was much flatter, and thus much less discriminating, than the Rasch theoretical IRF. Therefore, the extra flattening of the observed IRF introduced by using a higher bandwidth parameter did not wash out the systematic departure of the observed IRF from the theoretical IRF. This explains why, for item 46, power did not decline with large amounts of smoothing.

Figure 32 shows smoothed IRFs for item 47, which was a 2PL item with a discrimination of 2 instead of the Rasch implied discrimination of 1. As always, the IRF when the bandwidth was 0.05 was much more jagged than the generating model would suggest. With a bandwidth of 0.2, most of this jaggedness was smoothed out, but the IRF still retained its high level of discrimination. While the IRF for 0.8 bandwidth was quite smooth, it was also noticeably flattened, like with item 46. However, this item was supposed to appear highly discriminating. The flattening effect introduced by the large bandwidth parameter essentially flattened the IRF to almost perfectly resemble a theoretical Rasch IRF. This figure clearly shows why item 47 followed the “happy medium” pattern and also why power was 0 when large bandwidths were used.

Figure 33 shows smoothed IRFs for item 48, which was the 4PL item with an upper asymptote below 1. Figure 33 does not appear to explain the power results found for this item. Item 48 followed the “happy medium” pattern, which meant power was low for the large bandwidth parameters. However, as shown in Figure 33, the IRF when using a bandwidth of 0.8 still contains the non-1 upper asymptote specified by the 4PL model. Thus, it still contains the systematic departure of the observed IRF from the theoretical
IRF. Item 48 misfit in two ways: higher discrimination and an upper asymptote. When smoothed, the presence of the upper asymptote caused to the IRS to flatten and be closer to the average item discrimination. Although the upper asymptote was still visible with smoothing, relatively few examinees were in the ability range most impacted by the upper asymptote. Thus, with large amounts of smoothing, this item appeared to fit well.

Figure 34 shows smoothed IRFs for item 49, which was generated using a hyperbolic cosine model. The power results for this item followed the plateau pattern. Figure 34 helps confirm why item 49 followed this pattern. As can be seen, the IRF using a bandwidth of 0.2 still retained the large systematic departure from the theoretical IRF at the high end of the ability scale, while also smoothing out much of the jaggedness. While this systematic departure was less extreme when the bandwidth was 0.8, it is still very clearly present. This helps explain why power did not decrease as the bandwidth increased, leading to the “plateau” pattern for item 49.

Lastly, Figure 35 shows smoothed IRFs for item 50, which was the “flat middle” item. Like item 49, item 50 followed the “plateau” pattern. Also like item 49, the observed IRF for item 50 when the bandwidth was 0.8 still clearly preserved the flatness of the curve in the middle section of the ability distribution. Visually, it appeared to preserve the flatness better than when the bandwidth was 0.2, given the relative jaggedness still present in the IRF. Thus, it makes sense that large amounts of smoothing did not yield a decrease in power.

Research Question 3

As shown in Tables 1 and 2, type 1 error rates for Hanning and Kernel RISE were both around the nominal level of 0.05. Thus, smoothing technique did not have an effect
on the type 1 error rates of RISE. In terms of power, Hanning and Kernel smoothing did not always yield equivalent results, or patterns of results. For example, item 44 followed the “no smoothing” pattern when using Hanning smoothing, but followed the “happy medium” pattern when using Kernel smoothing. Similarly, item 50 followed the “happy medium” pattern when using Hanning and the “plateau” pattern when using Kernel smoothing. Additionally, the peak power levels of Hanning RISE and Kernel RISE were not identical for many item by sample size combinations. However, as Figures 22 through 28 show, peak power results for Hanning and Kernel RISE were generally quite similar. The only item where the peak power result of Hanning RISE and Kernel RISE was drastically different was item 47, the 2PL item. For this item, Hanning RISE performed much better than Kernel RISE. For the rest of the items, when there were differences between the two RISE statistics, they were most pronounced at a sample size of 200. In these cases, Kernel RISE outperformed Hanning RISE. It would be up to the practitioner to decide whether the differences in power between Kernel RISE and Hanning RISE at low sample sizes were enough to justify a choice of Kernel RISE over Hanning RISE. At the larger sample sizes, except for item 47, the choice of smoothing technique was relatively inconsequential, in terms of power.

**Research Question 4**

As can be seen in Figures 22-28, Outfit generally performed the best of the fit statistics, in terms of power, across the seven items. Infit tended to perform the second best, although its power for items 45 and 48 was much lower than any of the three other statistics. Hanning and Kernel RISE performed similarly to each other and worse than Infit and Outfit across the seven items. There were only three of the twenty-one item by
sample size encounters where either RISE statistic outperformed both Infit and Outfit: Kernel RISE for item 44 with a sample size of 200, and both Kernel and Hanning RISE for item 45 when the sample size was 200 or 500.

For all other item by sample size encounters, either Infit, Outfit, or both, had equal or better power than either RISE statistic. In many cases, the RISE statistics performed similarly to Infit, Outfit, or both, especially when the sample size was 1,000. However, they very rarely were sensitive to a type of misfit that either Infit or Outfit were not sensitive to. In fact, it was Infit that was most different in comparison to the other statistics, in terms of the types of misfit it was sensitive to. For example, as stated earlier, Infit had almost no power to detect the misfit of item 45 or 48, whereas the other three statistics were sensitive to these types of misfit. On the other hand, Outfit tended to be sensitive to the same types of misfit as both RISE statistics. But, within each type of misfit, Outfit was almost always equal or superior to the RISE statistics. This does not mean that the power of Hanning and Kernel RISE were bad. As can be seen in Figures 22-28, their power was often quite good, especially at higher sample sizes. They just generally were not as good as Outfit, nor did they pick up on types of misfit that Outfit did not also pick up on.

Differences amongst the four statistics were most prevalent when the sample size was 200. For some items, when the sample size was 500 or especially 1,000, all four statistics were so highly powered that choosing between them would be inconsequential. But, when there were differences between the statistics, it was Outfit that generally was the top performer.
Limitations

The research questions addressed in this study could not be addressed with real data. Thus, this study was a simulation study. There are inherent benefits to using simulated data. Because the data was simulated, it was known which items fit and did not fit the Rasch model. Additionally, for the items that did not fit the Rasch model, the nature and degree of the misfit was also known. Because many replications of each condition could be conducted, type 1 error rates and power were easy to calculate and examine.

However, simulation studies also bring limitations. For example, only seven types of misfit were examined in this study. Real data may misfit the Rasch model in ways not simulated in this study. Or, real data may misfit the Rasch model in similar ways as this study, but to greater or lesser degrees. For example, item 46 followed a 3PL model with a c-parameter of .2 and a true difficulty of 1. Perhaps if the c-parameter was .1 instead, or if the item had a different difficulty, comparisons of power across the four fit statistics may have been different.

In a similar vein, only two types of smoothing techniques were used in this study: Hanning and Kernel smoothing with a Gaussian function. There are many other methods for smoothing data, such as log-linear smoothing or cubic spline smoothing. If these types of smoothing had been used instead, perhaps RISE would have performed differently than observed in this study. Smoothing the raw item responses, instead of the proportion correct at each ability level (as was done in this study), could have also led to differing results than those found in this study.
Nine smoothing amounts were used and relatively clear patterns emerged for each item in terms of the relationship between smoothing and power. For the majority of these items, power tended to peak at a certain smoothing amount. For example, of the four items that followed the “happy medium” pattern for Kernel smoothing, power for three of those items peaked with a bandwidth of 0.2. However, that does not mean necessarily that 0.2 was the optimal bandwidth for those items. For example, perhaps power for item 45 would have technically peaked at a bandwidth of 0.184 instead of 0.2. There are an infinite number of bandwidth parameters that were not included in this study.

Finally, conditions were only replicated 500 times. Thus, the results could contain at least a small amount of sampling error, which can belie some of the observed differences in power of the fit statistics. For example, for item 44, power for Hanning RISE was 0.076 and 0.074 for Outfit. Differences this small between the two statistics may have been due simply to sampling. If more replications had been conducted, then there could have been more confidence that differences this small still constituted true differences, as opposed to differences due to sampling error. However, differences as small as the previous example would essentially be meaningless.

**Recommendations**

Given the limitations stated above, it would be unwise to offer strict recommendations that generalize to all possible scenarios where the Rasch model is used. As with any simulation study, caution must be heeded when interpreting and using the results. However, there were patterns that emerged from the study. These patterns may be used to at least offer practitioners some guidance when assessing the fit of their data to the Rasch model.
In terms of choosing a fit statistic for the Rasch model, it may only be necessary to examine Infit and Outfit. They sometimes were differentially sensitive to certain types of misfit. Infit in particular had a few items where its power was extremely poor, and a few items where it was the top performer. However, when used in tandem with Outfit, as is typically done with Rasch modeling, they equaled or outperformed RISE for almost every sample size and item combination. RISE, despite the fact that it is calculated and interpreted differently than Infit and Outfit, was not uniquely sensitive to certain types of misfit, in comparison to the combination of Infit and Outfit. Further, Infit and Outfit are currently much more readily accessible. Both Infit and Outfit are automatically provided by many Rasch software packages. RISE, on the other hand, must be computed manually. This is a much more time-consuming process, given the need for bootstrapping to compute the empirical null sampling distribution of RISE. Thus, given that Infit and Outfit are both simpler to obtain and perform better than RISE, a Rasch practitioner may find it unnecessary to calculate RISE and instead just examine Infit and Outfit.

If a Rasch practitioner still wants to use RISE, then this study provides some guidance on the choice of smoothing technique and amount. In terms of smoothing technique, the results were generally similar when using Hanning or Kernel smoothing. Thus, choice between these two techniques does not seem particularly important. However, it must be noted that there may be other smoothing techniques, ones that were not explored in this study, that would perform better.

Unlike smoothing technique, smoothing amount would generally be an important decision for a Rasch practitioner. As shown in this study, there was often a strong relationship between smoothing amount and power. Based on the results of this study, a
bandwidth around 0.2 seems like the most advisable choice for Kernel smoothing. Similarly, 5 smoothing iterations appears to be the best choice for Hanning smoothing.

This is because when power did reach a clear peak, it tended to peak at these values. For some items, power never reached a clear peak, and instead simply plateaued across a wide range of smoothing amounts. However, in these cases, a bandwidth of 0.2 or 5 smoothing iterations still performed comparably well to the higher smoothing amounts.

Yet, in cases when power did peak at a bandwidth of 0.2 or 5 smoothing iterations, power was often much worse for higher amounts of smoothing. Thus, a bandwidth of 0.2 or 5 smoothing iterations were generally either the best or nearly the best amounts of smoothing across all 7 items. Thus, they seem like the safest choices, based on the results of this study.

**Future Research**

There are many opportunities for future research to build upon this study. For example, this study only examined fit as it pertains to the Rasch model. Future studies could examine the performance of RISE when assessing the fit of data to the 2PL or 3PL model. This has been done in some previous studies, but not as it pertains to smoothing amount. Perhaps the relationship between smoothing amount and power would be different when examining the fit of data to a more complex IRT model.

Additionally, future research could examine different types of smoothing techniques than those used in this study. As stated in the limitations section, there are many techniques for smoothing data. The power of RISE may differ if these techniques were used.
New research could also compare Infit and Outfit to RISE, but with different types or degrees of misfit than those used in this study. This study found that RISE was not differentially sensitive to certain types of misfit than the tandem of Infit and Outfit. However, the types of misfit used in this study certainly do not exhaust all possible types of misfit to the Rasch model. Future researchers may find that RISE does in fact outperform Infit and Outfit for certain types of misfit not modeled in this study.
### Table 1

**Type 1 Error Rates for Hanning RISE**

<table>
<thead>
<tr>
<th>Smoothing Iterations</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>0</td>
<td>0.055</td>
</tr>
<tr>
<td>5</td>
<td>0.053</td>
</tr>
<tr>
<td>10</td>
<td>0.053</td>
</tr>
<tr>
<td>15</td>
<td>0.052</td>
</tr>
<tr>
<td>20</td>
<td>0.052</td>
</tr>
<tr>
<td>25</td>
<td>0.051</td>
</tr>
<tr>
<td>30</td>
<td>0.051</td>
</tr>
<tr>
<td>35</td>
<td>0.050</td>
</tr>
<tr>
<td>40</td>
<td>0.050</td>
</tr>
</tbody>
</table>

### Table 2

**Type 1 Error Rates for Kernel RISE**

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>0.05</td>
<td>0.053</td>
</tr>
<tr>
<td>0.1</td>
<td>0.053</td>
</tr>
<tr>
<td>0.2</td>
<td>0.051</td>
</tr>
<tr>
<td>0.3</td>
<td>0.053</td>
</tr>
<tr>
<td>0.4</td>
<td>0.052</td>
</tr>
<tr>
<td>0.5</td>
<td>0.053</td>
</tr>
<tr>
<td>0.6</td>
<td>0.055</td>
</tr>
<tr>
<td>0.7</td>
<td>0.055</td>
</tr>
<tr>
<td>0.8</td>
<td>0.054</td>
</tr>
</tbody>
</table>

### Table 3

**Type 1 Error Rates for Infit and Outfit**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>Infit</td>
<td>0.017</td>
</tr>
<tr>
<td>Outfit</td>
<td>0.027</td>
</tr>
</tbody>
</table>
Table 4

*Power of Infit*

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Item</th>
<th>200</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>0.046</td>
<td>0.186</td>
<td>0.292</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>0.012</td>
<td>0.008</td>
<td>0.028</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>0.800</td>
<td>0.990</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>0.800</td>
<td>0.990</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>0.068</td>
<td>0.118</td>
<td>0.194</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>0.796</td>
<td>0.994</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.758</td>
<td>0.992</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Table 5

*Power of Outfit*

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Item</th>
<th>200</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>0.074</td>
<td>0.234</td>
<td>0.372</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>0.752</td>
<td>0.976</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>0.800</td>
<td>0.990</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>0.684</td>
<td>0.984</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>0.448</td>
<td>0.746</td>
<td>0.950</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>0.960</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.422</td>
<td>0.854</td>
<td>0.992</td>
<td></td>
</tr>
</tbody>
</table>

Table 6

*Power of Hanning RISE on Item 44: Wavy*

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Smoothing Iterations</th>
<th>200</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.076</td>
<td>0.176</td>
<td>0.296</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.066</td>
<td>0.156</td>
<td>0.216</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.050</td>
<td>0.082</td>
<td>0.102</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.040</td>
<td>0.052</td>
<td>0.070</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.032</td>
<td>0.046</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.032</td>
<td>0.044</td>
<td>0.044</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.030</td>
<td>0.050</td>
<td>0.038</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>0.032</td>
<td>0.046</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0.030</td>
<td>0.042</td>
<td>0.038</td>
<td></td>
</tr>
</tbody>
</table>
Table 7
Power of Hanning RISE on Item 45: Big Dip

<table>
<thead>
<tr>
<th>Smoothing Iterations</th>
<th>Sample Size 200</th>
<th>Sample Size 500</th>
<th>Sample Size 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.558</td>
<td>0.988</td>
<td>1.000</td>
</tr>
<tr>
<td>5</td>
<td>0.856</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>10</td>
<td>0.822</td>
<td>0.998</td>
<td>1.000</td>
</tr>
<tr>
<td>15</td>
<td>0.78</td>
<td>0.994</td>
<td>0.998</td>
</tr>
<tr>
<td>20</td>
<td>0.728</td>
<td>0.996</td>
<td>1.000</td>
</tr>
<tr>
<td>25</td>
<td>0.632</td>
<td>0.978</td>
<td>1.000</td>
</tr>
<tr>
<td>30</td>
<td>0.584</td>
<td>0.956</td>
<td>0.996</td>
</tr>
<tr>
<td>35</td>
<td>0.480</td>
<td>0.918</td>
<td>0.998</td>
</tr>
<tr>
<td>40</td>
<td>0.414</td>
<td>0.838</td>
<td>0.998</td>
</tr>
</tbody>
</table>

Table 8
Power of Hanning RISE on Item 46: 3PL

<table>
<thead>
<tr>
<th>Smoothing Iterations</th>
<th>Sample Size 200</th>
<th>Sample Size 500</th>
<th>Sample Size 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.358</td>
<td>0.674</td>
<td>0.942</td>
</tr>
<tr>
<td>5</td>
<td>0.468</td>
<td>0.848</td>
<td>0.982</td>
</tr>
<tr>
<td>10</td>
<td>0.500</td>
<td>0.862</td>
<td>0.990</td>
</tr>
<tr>
<td>15</td>
<td>0.502</td>
<td>0.866</td>
<td>0.986</td>
</tr>
<tr>
<td>20</td>
<td>0.506</td>
<td>0.868</td>
<td>0.982</td>
</tr>
<tr>
<td>25</td>
<td>0.500</td>
<td>0.866</td>
<td>0.980</td>
</tr>
<tr>
<td>30</td>
<td>0.506</td>
<td>0.858</td>
<td>0.984</td>
</tr>
<tr>
<td>35</td>
<td>0.496</td>
<td>0.844</td>
<td>0.966</td>
</tr>
<tr>
<td>40</td>
<td>0.506</td>
<td>0.846</td>
<td>0.982</td>
</tr>
</tbody>
</table>
Table 9
*Power of Hanning RISE on Item 47: 2PL*

<table>
<thead>
<tr>
<th>Smoothing Iterations</th>
<th>Sample Size</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
<td>500</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.002</td>
<td>0.056</td>
<td>0.324</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.108</td>
<td>0.526</td>
<td>0.922</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.126</td>
<td>0.508</td>
<td>0.914</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.102</td>
<td>0.434</td>
<td>0.858</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.078</td>
<td>0.358</td>
<td>0.772</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.054</td>
<td>0.260</td>
<td>0.616</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.030</td>
<td>0.176</td>
<td>0.486</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>0.020</td>
<td>0.090</td>
<td>0.374</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0.006</td>
<td>0.044</td>
<td>0.168</td>
<td></td>
</tr>
</tbody>
</table>

Table 10
*Power of Hanning RISE on Item 48: 4PL*

<table>
<thead>
<tr>
<th>Smoothing Iterations</th>
<th>Sample Size</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
<td>500</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.138</td>
<td>0.270</td>
<td>0.634</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.238</td>
<td>0.552</td>
<td>0.900</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.232</td>
<td>0.518</td>
<td>0.860</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.198</td>
<td>0.450</td>
<td>0.786</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.186</td>
<td>0.360</td>
<td>0.730</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.164</td>
<td>0.326</td>
<td>0.626</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.148</td>
<td>0.280</td>
<td>0.530</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>0.138</td>
<td>0.238</td>
<td>0.514</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0.122</td>
<td>0.220</td>
<td>0.384</td>
<td></td>
</tr>
</tbody>
</table>
Table 11

*Power of Hanning RISE on Item 49: Hyperbolic Cosine Model*

<table>
<thead>
<tr>
<th>Smoothing Iterations</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>0</td>
<td>0.660</td>
</tr>
<tr>
<td>5</td>
<td>0.752</td>
</tr>
<tr>
<td>10</td>
<td>0.754</td>
</tr>
<tr>
<td>15</td>
<td>0.756</td>
</tr>
<tr>
<td>20</td>
<td>0.744</td>
</tr>
<tr>
<td>25</td>
<td>0.748</td>
</tr>
<tr>
<td>30</td>
<td>0.746</td>
</tr>
<tr>
<td>35</td>
<td>0.744</td>
</tr>
<tr>
<td>40</td>
<td>0.734</td>
</tr>
</tbody>
</table>

Table 12

*Power of Hanning RISE on item 50: Flat Middle*

<table>
<thead>
<tr>
<th>Smoothing Iterations</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>0</td>
<td>0.408</td>
</tr>
<tr>
<td>5</td>
<td>0.614</td>
</tr>
<tr>
<td>10</td>
<td>0.610</td>
</tr>
<tr>
<td>15</td>
<td>0.604</td>
</tr>
<tr>
<td>20</td>
<td>0.578</td>
</tr>
<tr>
<td>25</td>
<td>0.518</td>
</tr>
<tr>
<td>30</td>
<td>0.504</td>
</tr>
<tr>
<td>35</td>
<td>0.464</td>
</tr>
<tr>
<td>40</td>
<td>0.432</td>
</tr>
</tbody>
</table>
Table 13
*Power of Kernel RISE on Item 44: Wavy*

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>Sample Size 200</th>
<th>Sample Size 500</th>
<th>Sample Size 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.074</td>
<td>0.198</td>
<td>0.320</td>
</tr>
<tr>
<td>0.1</td>
<td>0.086</td>
<td>0.222</td>
<td>0.354</td>
</tr>
<tr>
<td>0.2</td>
<td>0.076</td>
<td>0.180</td>
<td>0.238</td>
</tr>
<tr>
<td>0.3</td>
<td>0.062</td>
<td>0.116</td>
<td>0.126</td>
</tr>
<tr>
<td>0.4</td>
<td>0.040</td>
<td>0.100</td>
<td>0.074</td>
</tr>
<tr>
<td>0.5</td>
<td>0.044</td>
<td>-</td>
<td>0.060</td>
</tr>
<tr>
<td>0.6</td>
<td>0.042</td>
<td>0.078</td>
<td>0.062</td>
</tr>
<tr>
<td>0.7</td>
<td>0.040</td>
<td>0.084</td>
<td>0.056</td>
</tr>
<tr>
<td>0.8</td>
<td>0.042</td>
<td>0.074</td>
<td>0.060</td>
</tr>
</tbody>
</table>

Table 14
*Power of Kernel RISE on Item 45: Big Dip*

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>Sample Size 200</th>
<th>Sample Size 500</th>
<th>Sample Size 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.614</td>
<td>0.990</td>
<td>1.000</td>
</tr>
<tr>
<td>0.1</td>
<td>0.790</td>
<td>1.000</td>
<td>0.998</td>
</tr>
<tr>
<td>0.2</td>
<td>0.836</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.3</td>
<td>0.822</td>
<td>0.998</td>
<td>1.000</td>
</tr>
<tr>
<td>0.4</td>
<td>0.712</td>
<td>0.986</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>0.540</td>
<td>-</td>
<td>0.994</td>
</tr>
<tr>
<td>0.6</td>
<td>0.376</td>
<td>0.696</td>
<td>0.902</td>
</tr>
<tr>
<td>0.7</td>
<td>0.262</td>
<td>0.464</td>
<td>0.684</td>
</tr>
<tr>
<td>0.8</td>
<td>0.192</td>
<td>0.308</td>
<td>0.454</td>
</tr>
</tbody>
</table>
Table 15

*Power of Kernel RISE on Item 46: 3PL*

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>Sample Size 200</th>
<th>Sample Size 500</th>
<th>Sample Size 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.370</td>
<td>0.698</td>
<td>0.960</td>
</tr>
<tr>
<td>0.1</td>
<td>0.474</td>
<td>0.834</td>
<td>0.976</td>
</tr>
<tr>
<td>0.2</td>
<td>0.538</td>
<td>0.878</td>
<td>0.982</td>
</tr>
<tr>
<td>0.3</td>
<td>0.564</td>
<td>0.888</td>
<td>0.986</td>
</tr>
<tr>
<td>0.4</td>
<td>0.584</td>
<td>0.902</td>
<td>0.996</td>
</tr>
<tr>
<td>0.5</td>
<td>0.602</td>
<td>-</td>
<td>0.996</td>
</tr>
<tr>
<td>0.6</td>
<td>0.614</td>
<td>0.900</td>
<td>0.994</td>
</tr>
<tr>
<td>0.7</td>
<td>0.614</td>
<td>0.876</td>
<td>0.996</td>
</tr>
<tr>
<td>0.8</td>
<td>0.618</td>
<td>0.884</td>
<td>0.994</td>
</tr>
</tbody>
</table>

Table 16

*Power of Kernel RISE on Item 47: 2PL*

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>Sample Size 200</th>
<th>Sample Size 500</th>
<th>Sample Size 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.000</td>
<td>0.056</td>
<td>0.426</td>
</tr>
<tr>
<td>0.1</td>
<td>0.028</td>
<td>0.178</td>
<td>0.622</td>
</tr>
<tr>
<td>0.2</td>
<td>0.064</td>
<td>0.256</td>
<td>0.660</td>
</tr>
<tr>
<td>0.3</td>
<td>0.036</td>
<td>0.096</td>
<td>0.246</td>
</tr>
<tr>
<td>0.4</td>
<td>0.004</td>
<td>0.008</td>
<td>0.004</td>
</tr>
<tr>
<td>0.5</td>
<td>0.000</td>
<td>-</td>
<td>0.000</td>
</tr>
<tr>
<td>0.6</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.7</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.8</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Table 17

*Power of Kernel RISE on Item 48: 4PL*

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>0.05</td>
<td>0.156</td>
</tr>
<tr>
<td>0.1</td>
<td>0.196</td>
</tr>
<tr>
<td>0.2</td>
<td>0.248</td>
</tr>
<tr>
<td>0.3</td>
<td>0.220</td>
</tr>
<tr>
<td>0.4</td>
<td>0.188</td>
</tr>
<tr>
<td>0.5</td>
<td>0.134</td>
</tr>
<tr>
<td>0.6</td>
<td>0.094</td>
</tr>
<tr>
<td>0.7</td>
<td>0.086</td>
</tr>
<tr>
<td>0.8</td>
<td>0.078</td>
</tr>
</tbody>
</table>

Table 18

*Power of Kernel RISE on Item 49: Hyperbolic Cosine*

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>0.05</td>
<td>0.696</td>
</tr>
<tr>
<td>0.1</td>
<td>0.740</td>
</tr>
<tr>
<td>0.2</td>
<td>0.796</td>
</tr>
<tr>
<td>0.3</td>
<td>0.812</td>
</tr>
<tr>
<td>0.4</td>
<td>0.826</td>
</tr>
<tr>
<td>0.5</td>
<td>0.812</td>
</tr>
<tr>
<td>0.6</td>
<td>0.798</td>
</tr>
<tr>
<td>0.7</td>
<td>0.788</td>
</tr>
<tr>
<td>0.8</td>
<td>0.792</td>
</tr>
</tbody>
</table>
Table 19

*Power of Kernel RISE on Item 50: Flat Middle*

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>Sample Size</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
<td>500</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.430</td>
<td>0.890</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.570</td>
<td>0.942</td>
<td>0.998</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.668</td>
<td>0.972</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.702</td>
<td>0.964</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.704</td>
<td>0.966</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.698</td>
<td>-</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.694</td>
<td>0.964</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>0.666</td>
<td>0.962</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.632</td>
<td>0.952</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Table 20

*Comparing Peak Performances of Hanning and Kernel RISE to Infit and Outfit*

<table>
<thead>
<tr>
<th>Item</th>
<th>N</th>
<th>Infit</th>
<th>Outfit</th>
<th>Hanning RISE</th>
<th>Kernel RISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>200</td>
<td>0.046</td>
<td>0.074</td>
<td>0.076</td>
<td>0.086</td>
</tr>
<tr>
<td>44</td>
<td>500</td>
<td>0.186</td>
<td>0.234</td>
<td>0.176</td>
<td>0.222</td>
</tr>
<tr>
<td>44</td>
<td>1000</td>
<td>0.292</td>
<td>0.372</td>
<td>0.296</td>
<td>0.354</td>
</tr>
<tr>
<td>45</td>
<td>200</td>
<td>0.012</td>
<td>0.752</td>
<td>0.856</td>
<td>0.836</td>
</tr>
<tr>
<td>45</td>
<td>500</td>
<td>0.008</td>
<td>0.976</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>45</td>
<td>1000</td>
<td>0.028</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>46</td>
<td>200</td>
<td>0.800</td>
<td>0.800</td>
<td>0.506</td>
<td>0.618</td>
</tr>
<tr>
<td>46</td>
<td>500</td>
<td>0.990</td>
<td>0.990</td>
<td>0.868</td>
<td>0.902</td>
</tr>
<tr>
<td>46</td>
<td>1000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.990</td>
<td>0.996</td>
</tr>
<tr>
<td>47</td>
<td>200</td>
<td>0.800</td>
<td>1.000</td>
<td>0.922</td>
<td>0.660</td>
</tr>
<tr>
<td>47</td>
<td>500</td>
<td>0.990</td>
<td>0.984</td>
<td>0.526</td>
<td>0.256</td>
</tr>
<tr>
<td>47</td>
<td>1000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.922</td>
<td>0.660</td>
</tr>
<tr>
<td>48</td>
<td>200</td>
<td>0.068</td>
<td>0.448</td>
<td>0.238</td>
<td>0.248</td>
</tr>
<tr>
<td>48</td>
<td>500</td>
<td>0.118</td>
<td>0.746</td>
<td>0.552</td>
<td>0.534</td>
</tr>
<tr>
<td>48</td>
<td>1000</td>
<td>0.194</td>
<td>0.950</td>
<td>0.900</td>
<td>0.850</td>
</tr>
<tr>
<td>49</td>
<td>200</td>
<td>0.796</td>
<td>0.960</td>
<td>0.756</td>
<td>0.826</td>
</tr>
<tr>
<td>49</td>
<td>500</td>
<td>0.994</td>
<td>1.000</td>
<td>0.980</td>
<td>0.990</td>
</tr>
<tr>
<td>49</td>
<td>1000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>0.758</td>
<td>0.422</td>
<td>0.614</td>
<td>0.704</td>
</tr>
<tr>
<td>50</td>
<td>500</td>
<td>0.992</td>
<td>0.854</td>
<td>0.964</td>
<td>0.972</td>
</tr>
<tr>
<td>50</td>
<td>1000</td>
<td>1.000</td>
<td>0.992</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Figure 1. Comparison of the Rasch implied IRF and the generating IRF for item 44.
Figure 2. Comparison of the Rasch implied IRF and the generating IRF for item 45.
Figure 3. Comparison of the Rasch implied IRF and the generating IRF for item 46.
Figure 4. Comparison of the Rasch implied IRF and the generating IRF for item 47.
Figure 5. Comparison of the Rasch implied IRF and the generating IRF for item 48.
Figure 6. Comparison of the Rasch implied IRF and the generating IRF for item 49.
Figure 7. Comparison of the Rasch implied IRF and the generating IRF for item 50.
Figure 8. Power results for Hanning RISE on Item 44, the wavy item.
Figure 9. Power results for Hanning RISE on Item 45, the big dip item.
Figure 10. Power results for Hanning RISE on Item 46, the 3PL item.
Figure 11. Power results for Hanning RISE on Item 47, the 2PL item.
Figure 12. Power results for Hanning RISE on Item 48, the 4PL item.
Figure 13. Power results for Hanning RISE on Item 49, the hyperbolic cosine item.
Figure 14. Power results for Hanning RISE on Item 50, the flat middle item.
Figure 15. Power results for Kernel RISE on Item 44, the wavy item.
Figure 16. Power results for Kernel RISE on Item 45, the big dip item.
Figure 17. Power results for Kernel RISE on Item 46, the 3PL item.
Figure 18. Power results for Kernel RISE on Item 47, the 2PL item.
Figure 19. Power results for Kernel RISE on Item 48, the 4PL item.
Figure 20. Power results for Kernel RISE on Item 49, the hyperbolic cosine item.
Figure 21. Power results for Kernel RISE on Item 50, the flat middle item.
Figure 22. Comparing the power of Infit, Outfit, Peak Hanning RISE, and Peak Kernel RISE for item 44, the wavy item.
Figure 23. Comparing the power of Infit, Outfit, Peak Hanning RISE, and Peak Kernel RISE for item 45, the big-dip item.
Figure 24. Comparing the power of Infit, Outfit, Peak Hanning RISE, and Peak Kernel RISE for item 46, the 3PL item.
Figure 25. Comparing the power of Infit, Outfit, Peak Hanning RISE, and Peak Kernel RISE for item 47, the 2PL item.
Figure 26. Comparing the power of Infit, Outfit, Peak Hanning RISE, and Peak Kernel RISE for item 48, the 4PL item.
Figure 27. Comparing the power of Infit, Outfit, Peak Hanning RISE, and Peak Kernel RISE for item 49, the hyperbolic cosine item.
Figure 28. Comparing the power of Infit, Outfit, Peak Hanning RISE, and Peak Kernel RISE for item 50, the flat middle item.
Figure 29. Illustration of the effects of different bandwidth parameters on the item response function for item 44, the way item.
Figure 30. Illustration of the effects of different bandwidth parameters on the item response function for item 45, the big dip item.
Figure 31. Illustration of the effects of different bandwidth parameters on the item response function for item 46, the 3PL item.
Figure 32. Illustration of the effects of different bandwidth parameters on the item response function for item 47, the 2PL item.
Figure 33. Illustration of the effects of different bandwidth parameters on the item response function for item 48, the 4PL item.
Figure 34. Illustration of the effects of different bandwidth parameters on the item response function for item 49, the hyperbolic cosine item.
Figure 35. Illustration of the effects of different bandwidth parameters on the item response function for item 50, the flat middle item.
References


