Local Item Response Theory for Detection of Spatially Varying Differential Item Functioning

Samantha Robinson

University of Arkansas, Fayetteville

Follow this and additional works at: http://scholarworks.uark.edu/etd

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation

http://scholarworks.uark.edu/etd/2520

This Dissertation is brought to you for free and open access by ScholarWorks@UARK. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of ScholarWorks@UARK. For more information, please contact scholar@uark.edu, ccmiddle@uark.edu.
Local Item Response Theory for Detection of Spatially Varying Differential Item Functioning

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Educational Statistics and Research Methods

by

Samantha Robinson
University of Arkansas
Bachelor of Arts in Mathematics, 2009
University of Arkansas
Master of Science in Statistics, 2012

December 2017
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

Dr. Ronna Turner
Dissertation Director

Dr. Wen-Juo Lo
Committee Member

Dr. Andronikos Mauromoustakos
Committee Member
Abstract

Mappings of spatially-varying Item Response Theory (IRT) parameters are proposed, allowing for visual investigation of potential Differential Item Functioning (DIF) based upon geographical location without need for pre-specified groupings and before any confirmatory DIF testing. This proposed model is a localized approach to IRT modeling and DIF detection that provides a flexible framework, with current emphasis being on 1PL/Rasch and 2PL models. Applications to both simulated and empirical survey data, utilizing a box-car kernel weighting scheme with several fixed bandwidths on irregular spatial lattices, are presented both to demonstrate the methodology and to illustrate the benefit of localized IRT modeling. There is not only practical value with this method but also visual appeal when initial attempts to consider measurement invariance are being made across national, state, or other political and geographical boundaries, especially when comparisons are made to traditional DIF techniques. This approach, making use of surface mappings of estimated item parameters, serves to detect DIF across space without a priori groupings, thereby identifying regional disparities and latent spatial trends in item functionality that may be unobservable on a more aggregate, global level.
Acknowledgments

I would like to thank the Educational Statistics and Research Methods (ESRM) faculty and students. In particular, I would like to thank Dr. Wen-Juo Lo and Dr. Ronna Turner. Without their guidance and support, this work would have never come to fruition.

Dr. Wen-Juo Lo provided encouragement and reassurance. Dr. Ronna Turner provided the same and, also, an unwavering confidence in me and in this work that I will never be able to express my gratitude for properly.

I would also like to thank Dr. Andronikos Mauromoustakos. His perspective and his comments have not only shaped this current work but have also provided a considerable list of future directions for its continuation.

In addition, I thank the UA Department of Mathematics. This department has continued to put faith in my teaching ability for nearly seven years and, by letting me hold a position, has provided for me the means to continue my education and to complete this work.

Finally, I am thankful to Giovanni for the coffee. I am thankful to Ian for the walks. To all my friends and my family, I am thankful for the support that you provided me during these years. You may not have even known how much your actions, large or small, mattered during this time and you may never know how honored I am to have you in my life.
Dedication

To my Bella Rose. This work alone, though not nearly enough, will be the first of so many things in my life dedicated to you.
# Table of Contents

Introduction ................................................................................................................................. 1

Motivation of Study ...................................................................................................................... 8

Background and Need of Study .................................................................................................. 10

Purpose of Study ......................................................................................................................... 12

Significance of Study .................................................................................................................. 12

Literature Review ....................................................................................................................... 16

Overview of Spatial Data Analysis ............................................................................................. 16

Local Spatial Data Analysis ......................................................................................................... 26

Geographically Weighted Regression ......................................................................................... 34

Item Response Theory ................................................................................................................ 42

Differential Item Functioning ...................................................................................................... 52

Method ........................................................................................................................................ 61

Local Item Response Theory ...................................................................................................... 61

Procedural Steps for the Proposed Method ................................................................................ 68

Simulated Data ............................................................................................................................ 70

Empirical Data ............................................................................................................................. 77

Case Study Descriptions ............................................................................................................ 79

Results ......................................................................................................................................... 90

Case Study 1 – Simulated Data Exhibiting No Regional DIF ....................................................... 90

Case Study 2 – Simulated Data Exhibiting Regional DIF with Spatial Clustering .......... 110

Case Study 3 – Simulated Data Exhibiting Regional DIF with No Spatial Clustering ...... 126
List of Tables

Table 1 Kernel functions available for GW modeling ......................................................... 39

Table 2 Mantel-Haenszel contingency table conditioned on the \( j \)th raw test score .............. 57

Table 3 The Malawi DHS complete family planning knowledge instrument items .................. 79

Table 4 Case study descriptions of non-manipulated and manipulated factors ....................... 89

Table 5 Assigned county sample sizes .................................................................................. 91

Table 6 1PL local IRT model summary – CS 1 estimated difficulty coefficients averaged across counties ................................................................................................................................. 93

Table 7 1PL local IRT model summary – CS 1 estimated difficulty coefficients for question 1 by county .................................................................................................................................................. 95

Table 8 1PL local IRT model summary – CS 1 standard errors for estimated difficulty coefficients for question 1 by county ........................................................................................................................................... 96

Table 9 2PL local IRT model summary – CS 1 estimated discrimination coefficients averaged across counties ................................................................................................................................................. 98

Table 10 2PL local IRT model summary – CS 1 estimated discrimination coefficients for question 1 by county .......................................................................................................................................................... 100

Table 11 2PL local IRT model summary – CS 1 standard errors for estimated discrimination coefficients for question 1 by county ............................................................................................................................................. 101

Table 12 Number of counties flagged for DIF on question 1 using pairwise comparisons .............. 106

Table 13 Local IRT model summary – CS 2 estimated difficulty and discrimination coefficients for question 1 across all 75 Arkansas counties ........................................................................................................... 112
Table 14 Local IRT model summary – CS 2 estimated difficulty and discrimination coefficients for question 1 across reference and focal Arkansas counties ................................................................. 113

Table 15 Traditional DIF methods flagging question 1 using a non-proximal approach in CS 2 ............................................. 118

Table 16 Local IRT model summary – CS 3 estimated difficulty and discrimination coefficients for question 1 across all 75 Arkansas counties ........................................................................ 129

Table 17 Local IRT model summary – CS 3 estimated difficulty and discrimination coefficients for question 1 across reference and focal Arkansas counties ........................................... 130

Table 18 Traditional DIF methods flagging question 1 using a non-proximal approach in CS 3 ................................................................. 135

Table 19 Local IRT model summary – CS 4 estimated difficulty coefficients for DHS items averaged across all 28 Malawi districts ........................................................................................................ 148
List of Figures

Figure 1 Dialect differences for rubber-soled shoes ................................................. 5
Figure 2 Dialect differences for sweetened carbonated beverages .............................. 6
Figure 3 Locations of Georgia county centroids and county seats ............................. 17
Figure 4 Locations of Chicago, Illinois neighborhood/district centroids ..................... 18
Figure 5 Neighborhood structure possibilities for regular lattice data ......................... 19
Figure 6 Neighborhood structure possibilities for Georgia counties .......................... 20
Figure 7 Neighborhood structure possibilities for Chicago neighborhoods/districts ......... 21
Figure 8 Mixture of point-referenced and irregular lattice spatial data for the London boroughs ................................................................. 21
Figure 9 Proximity/contiguity matrix weights ............................................................. 22
Figure 10 Illustration of discrete and fuzzy zoning systems ....................................... 23
Figure 11 Overlay grid of the London boroughs for regression point selection ............... 36
Figure 12 Item characteristic curves for 25 dichotomous items ................................... 46
Figure 13 Uniform differential item functioning favoring the reference group .............. 54
Figure 14 Non-uniform differential item functioning favoring the focal group at lower ability levels but favoring the reference group at higher ability levels ........................................ 55
Figure 15 Regular lattice with subgroups defined by cardinal direction ....................... 73
Figure 16 Regular lattice with two or three latent spatial subgroups ............................ 74
Figure 17 Irregular lattice of Arkansas counties for CS 1, 2, and 3 ............................... 80
Figure 18 Arkansas county relative enrollment sizes ................................................. 81
Figure 19 Arkansas metropolitan statistical areas for CS 2 ...................................... 82
Figure 20 Arkansas county subgroups defined by cardinal direction ........................... 84
Figure 21 Irregular lattice of Malawi districts for CS 4 ................................................................. 86
Figure 22 Malawi district subgroups defined by cardinal direction ........................................... 88
Figure 23 Neighborhood structures for Arkansas counties at three fixed bandwidths ............ 92
Figure 24 1PL local IRT model difficulty estimates and standard errors for CS 1 ...................... 94
Figure 25 2PL local IRT model discrimination estimates and standard errors for CS 1 ............ 99
Figure 26 1PL local IRT model local Moran’s I results for CS 1 ................................................ 104
Figure 27 2PL local IRT model local Moran’s I results for CS 1 ................................................ 105
Figure 28 1PL local IRT model spatial subgroups utilizing local Moran’s I for CS 1 ............... 107
Figure 29 2PL local IRT model spatial subgroups utilizing local Moran’s I for CS 1 ............... 109
Figure 30 CS 2 1PL local IRT model difficulty estimates across fixed bandwidths and three DIF magnitudes (small, moderate, large) .................................................................................. 114
Figure 31 CS 2 2PL local IRT model discrimination estimates across fixed bandwidths and DIF magnitudes (small, moderate, large) .................................................................................. 116
Figure 32 1PL local IRT model local Moran’s I results at three DIF magnitudes (small, moderate, large) for CS 2 ......................................................................................................................... 120
Figure 33 1PL local IRT model spatial subgroups utilizing local Moran’s I at three DIF magnitudes (small, moderate, large) for CS 2 ............................................................................................................... 121
Figure 34 2PL local IRT model local Moran’s I results at three DIF magnitudes (small, moderate, large) for CS 2 ......................................................................................................................... 122
Figure 35 2PL local IRT model spatial subgroups utilizing local Moran’s I at three DIF magnitudes (small, moderate, large) for CS 2 ............................................................................................................... 123
Figure 36 1PL local IRT model spatial subgroups utilizing local Moran’s I for mixed DIF CS 2 . . ........................................................................................................................................................................... 126
Figure 37 Arkansas reference and focal group assignments for CS 3 ........................................127

Figure 38 CS 3 1PL local IRT model difficulty estimates across fixed bandwidths and three DIF magnitudes (small, moderate, large) ..........................................................................................................................131

Figure 39 CS 3 2PL local IRT model discrimination estimates across fixed bandwidths and DIF magnitudes (small, moderate, large) .....................................................................................................................133

Figure 40 1PL local IRT model local Moran’s I results at three DIF magnitudes (small, moderate, large) for CS 3 ...........................................................................................................................................138

Figure 41 1PL local IRT model spatial subgroups utilizing local Moran’s I at three DIF magnitudes (small, moderate, large) for CS 3 ............................................................................................................................................139

Figure 42 2PL local IRT model local Moran’s I results at three DIF magnitudes (small, moderate, large) for CS 3 ..............................................................................................................................................140

Figure 43 2PL local IRT model spatial subgroups utilizing local Moran’s I at three DIF magnitudes (small, moderate, large) for CS 3 .............................................................................................................................................141

Figure 44 Neighborhood structures for Malawi districts at four fixed bandwidths .............145

Figure 45 Family planning method knowledge total score averages and standard deviations at the district level in Malawi ..............................................................................................................................................151

Figure 46 CS 4 1PL local IRT model difficulty estimates across fixed bandwidths for the DHS IUD item .........................................................................................................................................................154

Figure 47 CS 4 1PL local IRT model local Moran’s I results across fixed bandwidths for the DHS IUD item .........................................................................................................................................................155

Figure 48 1PL local IRT model spatial subgroups utilizing local Moran’s I at four fixed bandwidths for the DHS IUD item ..............................................................................................................................................156
Figure 49  CS 4 1PL local IRT model difficulty estimates across fixed bandwidths for the DHS implant item ............................................................... 158

Figure 50  CS 4 1PL local IRT model local Moran’s I results across fixed bandwidths for the DHS implant item .......................................................................................... 159

Figure 51  1PL local IRT model spatial subgroups utilizing local Moran’s I at four fixed bandwidths for the DHS implant item ................................................................. 160

Figure 52  Malawi urban and rural district classifications for CS 4 .......................................................... 161

Figure 53  Item characteristic curves for the DHS IUD and implant items by region, urban classification, and district ...................................................................................... 163
Chapter 1

Introduction

International large-scale assessments (ILSAs) are relatively recent endeavors created and implemented in the mid-1960s to compare educational achievement across nations (Hanushek & Woessmann, 2013). Developed in response to concern over an apparent inequitable distribution of human capital, ILSAs have grown in global importance (Kirsch, Lennon, von Davier, Gonzalez & Yamamoto, 2013). In fact, Braun (2013) argues that the steady increase in the number of participating countries demonstrates a recognition of and is a testament to the global importance of ILSAs. While not limited to educational assessments, examples of these types of ILSAs include surveys such as the Programme for International Student Assessment (PISA), International Adult Literacy Survey (IALS), the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the European Survey on Language Competences (ESLC) which are all used broadly to make cross-national comparisons of educational achievement.

The manner in which ILSAs shape the landscape of educational research is vast. According to Klieme (2013), ILSAs serve as indicators of educational system equity, provide knowledge about factors determining educational effectiveness, and necessarily create a comparative database to study questions of scientific and policy-oriented significance. For instance, these large-scale assessments provide a “monitoring structure” for educational systems (Klieme, 2013). ILSAs assist in investigating potential unbalance in human capital, believed to contribute both to the prosperity of a nation and to the quality of individual lives, and the impact any unbalance in human capital has on economic growth (Kirsch, Lennon, von Davier, Gonzalez & Yamamoto, 2013; Klieme, 2013). Moreover, ILSAs can function as change agents driving
reform, inasmuch as they provide transparency regarding educational systems (Braun, 2013; Ritzen, 2013).

Despite the understandable benefits of ILSAs, allowing for comparisons both within and across countries, the growing importance of the findings and the growing number of participating countries gives rise to a growing need to address inherent difficulties in test construction, adaptation, and score comparability. In particular, differential item functioning (DIF), typically seen as a threat to validity, is one difficulty arising in international surveys. Holland and Wainer define DIF as a relative term whereby an item performs differently for one group of examinees relative to the way it performs for another group of examinees (as cited in Zwitser, 2017). In fact, as noted by Zwitser (2017), items on international surveys are likely to exhibit DIF and prior research demonstrates that DIF exists in educational surveys.

Differences in item functionality can be attributed to any of several factors with those commonly investigated including class membership in gender, racial, ethnic, religious, or language subgroups (Apinyapibal, Lawthong, & Kanjanawasee, 2015; Tutz & Berger, 2016). To illustrate one difficulty arising from the multinational nature of ILSAs, consider the translation or adaptation of a survey instrument into multiple languages. While there are benefits to translating a well established instrument, including enhancing fairness by allowing examinees to test in a language of choice, the previously established reliability and validity of the instrument does not directly translate to the new language group and this process can create DIF (Hambleton & Kanjee, 1995). For instance, consider an example offered by Hambleton (1994, p. 235) where examinees are presented with the following item:

Where is a bird with webbed feet most likely to live?

a. In the mountains
b. In the woods
c. In the sea
In the desert

The above item was translated from English for Swedish-speaking examinees. Part of the item, in translation, becomes “swimming feet” rather than “webbed feet”. This translation gave Swedish-speaking examinees an understandable advantage on the item over their English-speaking counterparts. Any observable differences in group achievement based on this particular item should not necessarily be attributed to true differences in achievement but, rather, to the poor item resulting from the translation of the original item.

International surveys are occasionally limited in score comparability due to the occurrence of DIF. To avoid DIF arising from translation, research recommends the use of two, independent, bilingual translators familiar with the cultures of each group of examinees (Hambleton, 1994; Hambleton & Kanjee, 1995). Moreover, to minimize the risk of induced DIF, it is ideal for translators to understand the construct being measured, to be familiar with the subject matter, and to have some minimal training in test construction (Hambleton & Kanjee, 1995).

While it might seem apparent to exercise caution when translating or adapting international educational assessments, biases created due to underlying linguistic, psychological, social, or cultural differences can occur even when comparisons are being made within countries. Large scale assessments exist, also, on the national level and came into existence during roughly the same time period as ILSAs. For example, the National Assessment of Educational Progress (NAEP) was conducted first in the United States (US) in 1969 (Kirsch et al., 2013). Braun (2013) argues that, although cross-national comparisons are growing in importance and are of great interest, subnational comparisons, which are rarely given equal attention to their cross-national counterparts, have a greater immediate use. These subnational, within-country,
comparisons can be made with national assessments such as NAEP or with ILSAs, which likewise allow for subnational comparisons of student performance to be made (Klieme, 2013).

While caution is necessarily paid when dealing with the translation of international surveys, linguistic differences are not simply an issue of translation. Even minor linguistic differences within a particular language could impact the ability of a respondent on a particular item and these differences can occur within a country. Take for instance Figure 1 and Figure 2 adapted from Katz (2016) using data obtained from the Cambridge Online Survey of World Englishes data (Vaux, 2013).

While the two examples may not seem of concern in an educational setting, consider each example in a different survey context. For instance, the following is a hypothetical dichotomous item for a consumer behavior survey: *Will you buy a new pair of tennis shoes in the next three to six months? (Yes/No)*

The above item intends to assess customer demand for tennis shoes. Since one fundamental goal in merchandising is to market the right products in the right quantities to retailers and consumers, survey results from this particular consumer survey item could be used for merchandise distribution of tennis shoe products to retailers. If consumers in the Northeastern US appear to have lower levels of “demand” for tennis shoes, retailers in this region may receive fewer tennis shoe products in the upcoming months and, instead, may be sent alternative products (e.g., boots). However, any observable regional differences in “demand” for tennis shoes based on this particular item should not necessarily be attributed to true differences in “demand” but, rather, to lexical variation (see Figure 1). This above item would potentially exhibit regional DIF and the reason for the occurrence of DIF on this item could subsequently be investigated.
Consider, also, the following hypothetical dichotomous item for a consumer behavior survey intended to assess brand adherence: *Are you more likely to drink a coke with your meal than another available beverage? (Yes/No)*

The above item intends to assess customer brand adherence for the Coca-Cola brand but, unintentionally, is a poorly-worded question. Many consumers in the Southern US will appear to have higher levels of “brand adherence” for Coca-Cola. However, any observable regional differences in “brand adherence” for Coca-Cola based on this particular item should not necessarily be attributed to true differences in “Coca-Cola” adherence but, rather, to lexical variation (see Figure 2). This item would also potentially exhibit regional DIF and the reason for the occurrence of DIF on this item could subsequently be investigated.
There might also exist social and/or cultural differences within a particular country that could contribute to the differential functioning of a particular survey item. For instance, it is recognized by Cho and Gimpel (2010) that the patterns of certain political measures are uneven geographically and may vary by location. This variation might be the result of numerous factors not considered on an aggregate-level such as location-specific sociological factors in addition to historical and cultural forces (Cho & Gimpel, 2010). While any number of political constructs could form the basis for example, consider the measurement of a latent construct of political support for a particular candidate, Candidate X. At the same level of “political support for Candidate X”, the following hypothetical dichotomous item may function differently for respondents across geographic or spatial location: *Did you vote for Candidate X? (Yes/No)*
The above item intends to measure political support for Candidate X and seems a rather clearly-worded question. However, if the respondent actually votes for Candidate X but is located in a community that strongly opposes the candidate, the respondent might be less likely to answer the item in the affirmative than if they were located in a community that strongly supported the candidate. The social and cultural pressure of the community can impact the potential for the respondent to be deceptive on the item thereby suggesting that, at the same level of political support for Candidate X, the probability of answering the item in the affirmative differs based upon geographic or spatial location.

The above examples attempt to highlight the possibility of DIF due to geographic or spatial location, whether exploring cross-national or subnational comparisons of a measured latent trait. Besides educational assessments and the hypothetical consumer behavior items provided consider further applications to cross-national, survey-based marketing research. Several marketing studies have investigated the concept of Extreme Response Style (ERS), defined as a tendency of respondents on surveys to favor the endpoints of a rating scale independent of the specific item content (De Jong, Steenkamp, Fox, & Baumgartner, 2008). Research suggests that ERS, thought to contaminate rating scale data and distort the measurement of attitudes, is related to cultural orientations and is frequently viewed as a learned behavior, the result of socialization (De Jong et al., 2008; Peterson, Rhi-Perez, & Albaum, 2014). ERS is hypothesized to contribute to DIF in marketing surveys inasmuch as observed differences in a particular marketing construct may be interpreted as substantive differences when they are, in fact, attributable to country-specific variations in ERS and not in the marketing construct itself. Consequently, ERS has the potential to create regional DIF that, without additional investigation, could have adverse effects on the decisions of national and international marketers.
Differences in item functionality can be of concern in educational assessments, marketing surveys, medical screening tests (see Longford, 2014), personality inventories (see Huang, Church, & Katigbak, 1997), psychological instruments, and more. While the measured latent traits will be dissimilar for the variety of disciplines in which survey instruments are utilized, it is quite possible that items on these instruments may function differently across space due to regional and other geographic disparities.

**Motivation of Study**

Differential item functioning (DIF) occurs when items function differently for individuals of the same latent ability level based upon a class or group membership. To be concrete, the probability of answering an item successfully differs for individuals of the same latent ability level based upon a class or group membership (Zumbo, 1999). International Large-Scale Assessments (ILSAs), growing in importance and not limited only to cognitive ability tests and educational assessments, are designed with the intention to make cross-national and subnational comparisons (Hanushek & Woessmann, 2013). Despite careful test construction, items in this international context commonly exhibit DIF. It is also quite possible that items on educational assessments and other survey instruments may function differently due to geographic or spatial location, even within a nation. While the desire for comparable cross-national and cross-cultural comparisons has spurred the development of methods for detecting item-level DIF for many groups, representing the many participating countries in ILSAs, there is still a paucity of investigations into DIF arising on the subnational level.

Investigation of potential DIF based upon geographic or spatial location, whether national or subnational comparisons are to be made, is of great interest for a number of reasons. Firstly, observable regional disparities in item functionality might be directly attributable to differences
in the spatial location of observations; as such, the relationship between latent ability on a measured construct and responses on a particular item might exhibit some form of spatial nonstationarity and heterogeneity. In fact, Klieme (2013) hinted at the idea of spatial nonstationarity and heterogeneity in ILSAs by observing that, often, questionnaire scales show “strange behavior” when comparisons are made on differing levels such as comparisons made at the country level as opposed to comparisons made on the more local school level. This idea, in essence, suggests the idea of a spatial Simpson’s paradox whereby global, aggregated estimates of the relationships between latent ability and responses on particular items may obscure interesting geographical relationships that exist on a local level. Secondly, DIF is of serious concern in certain circumstances. It is seen as a threat to validity. It limits score comparability between groups. It can also lead to inappropriate decisions that have extremely adverse effects in high stakes contexts whether these contexts are educational, business, medical, or psychological in nature. Thirdly, these regional disparities in item functionality might only highlight the need for further investigation into alternate explanatory covariates, such as specific area teaching practices. Zwitser and Glaser (2017) emphasize that DIF on survey instruments can be viewed as an interesting outcome that need not invalidate other findings. Moreover, Hambleton and Kanjee (1995) remind us that DIF studies are invaluable but they are, by nature, statistical studies; the source of the problem(s) will not be identified without subsequent causal investigations. While DIF that is observable spatially might suggest an apparent unfair advantage for certain populations, any regional differences detected (e.g., hotspots or clusters of item functionality) could also be used to discover locations of academic excellence on certain items, concepts, or subscales. Consequently, applying this to an educational setting, areas exhibiting regional DIF may serve as valuable examples upon which educators may benchmark, thereby increasing the
educational opportunity available to all students regardless of geographic location. For the above reasons, the probable existence of regional DIF motivates the construction of a statistical tool that can identify regional differences in item functionality which occur across space.

**Background and Need of Study**

The idea of comparing groups that are location-specific is not new; in fact, utilizing location in some form as a grouping variable is now commonplace in educational measurement research due, in part, to the increasing predominance of ILSAs. However, traditional DIF detection methods require that only two groups be considered, the focal group and the reference group. As such, previous extensions of DIF detection methods to a multiple group setting to allow for comparison of groups that differ across space (e.g., countries) typically focus on either (1) naïve comparisons based on cardinal direction such as North versus South or East versus West or (2) the application of several pairwise comparisons of countries (Svetina & Rutkowski, 2014). As noted by Hambleton and Kanjee (1995), multiple pairwise comparisons prove to be time consuming and costly. The need to assess DIF simultaneously in all the groups or to reduce the number of pairwise comparisons conducted was the goal of many previous approaches such as that of Ellis and Kimmel (as cited in Hambleton & Kanjee, 1995). These approaches amounted to artificially creating two groups, one group against the aggregate of the other groups or against the composite of all groups. The Organisation for Economic Cooperation and Development (OECD) assessments of the PISA utilize an ANOVA-like procedure examining item-by-country interactions (as cited in Svetina & Rutkowski, 2014). Still other methods attempt to extend approaches of Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT) to a multiple group setting (Bock & Zimowski, 1997; Muthen & Christoffersson, 1981; Muthen & Lehman, 1985). Additional extensions and attempts to address the issue of DIF
detection in a multiple group setting include Multiple Indicator Multiple Cause (MIMIC) models, MIMIC-interaction models, DIF using a Lasso approach (LR Lasso DIF), the Alignment Method, and recursive partitioning approaches such as Rasch trees and item-focused trees (e.g., Berger & Tutz, 2016; Finch, 2005; Magis, Tuerlinckx & De Boeck, 2015; Muthen & Asparouhov, 2014; Strobl, Kopf & Zeileis, 2010; Tutz & Berger, 2016; Woods & Grimm, 2011). Many of the aforementioned methods still require a priori, pre-specified, grouping variables and do not consider interactions between these variables for group formation. As such, differences in item functionality that do not exist simply due to race/ethnicity, gender, socioeconomic status, country, or another pre-specified class or group membership, are not investigated. While some of these methods do allow for multiple factors, even quantitative continuous ‘factors’, and interactions of factors, the described results will be within the scope of pre-specified covariates (Apinyapibal et al., 2015). Moreover, none of the methods provided above take into consideration that country borders, even state/territory borders, are political borders and are, in many ways, seemingly arbitrary spatial boundaries. While these spatial groupings may be of interest for detecting DIF, the boundaries that dictate one state/territory from the next or one country from another are man-made and artificial. These political boundaries ignore potential differences that might exist, arising from differences in geography, landscape, language, bordering peoples, and more. Additionally, while multiple group DIF detection methods may adequately test measurement invariance across several groups, the conceivable spatial structure of the observations is not fully utilized or is, altogether, dismissed. Tobler (1970) observed what is known as the First Law of Geography, which assumes that near things are more related than distant things and underlies the concept of spatial autocorrelation. Any existing potential spatial autocorrelation and underlying spatial structure in observed data is
not currently exploited or used in the many available multiple group DIF detection methods. Consequently, there is a need to investigate potential regional DIF based upon geographic location without the need for pre-specified groupings and before any confirmatory DIF testing while also taking into consideration underlying spatial structure.

**Purpose of Study**

The purpose of the current study is twofold: (1) to propose a methodology for the examination of item-level regional DIF, motivated by the context of large-scale assessments where national and subnational comparisons are intended, based upon a localized approach to IRT modeling such that underlying spatial structure of observations is considered and (2) to describe and illustrate the methodology by providing detailed examples of several simulated case studies and one empirical application, with comparisons made to traditional DIF techniques.

**Significance of Study**

Investigation of DIF throughout a spatial region has typically focused upon one pairwise comparison based on cardinal direction or multiple pairwise comparisons based on arbitrary spatial boundaries such as political or geographical borders. The currently proposed methodology provides an exploratory analysis that can guide, in a data-driven way, subsequent analyses and provide a means to minimize the number of group comparisons in confirmatory multiple group DIF detection methods or reduce the many pairwise comparisons to one where two clearly defined spatial groups emerge and traditional techniques for DIF detection might be applied once focal and reference groups are identified.

Despite the approach of using multiple pairwise comparisons for DIF detection in a multiple group setting being both time consuming and costly, it may adequately test measurement invariance across several groups. However, previous approaches based upon the
idea of multiple pairwise comparisons and new multiple group DIF detection approaches rely on pre-specified groupings that can ignore potential differences in item functionality that arise from other causes. The currently proposed methodology adds to the growing literature on DIF detection by providing a statistical tool for the investigation of potential DIF based upon geographic location without the need for pre-specified groupings.

Certain recursive partitioning approaches to multiple group DIF detection provide the researcher with the ability to specify a priori several factors and, consequently, groupings are not inherently pre-specified though grouping will be within the scope of pre-specified covariates (Apinyapibal et al., 2015). However, these approaches, in line with all previously discussed approaches, still do not utilize the spatial structure of observations. The currently proposed methodology is a truly spatial technique for DIF detection. Spatial nonstationarity and heterogeneity is addressed directly by allowing estimated parameters to vary across space. Consequently, the spatial structure of observations is utilized and reliance upon political borders as proxy grouping variables for spatial location is no longer needed.

Using the idea that violations of IRT assumptions (such as parameter invariance) across identifiable spatial or regional groups provides us with a working definition of regional DIF in space, mappings of spatially-varying IRT parameters are also proposed, allowing for visual investigation of potential DIF based upon geographical location without need for pre-specified groupings and prior to any confirmatory DIF testing. This local approach to IRT modeling, including both 1PL and 2PL models, provides a flexible framework for regional DIF detection and is offered to expand the current methodology. Applications illustrate the benefit of localized IRT modeling as a pretesting method for questionnaire design, especially when comparisons are made to traditional DIF techniques. In addition, there is visual appeal when initial attempts to
consider measurement invariance are to be made across political boundaries. Making use of surface mappings of estimated parameters, the approach serves to detect DIF across space without a priori groupings, identifying regional disparities and latent spatial trends that may otherwise be unobservable. As such, the proposed localized IRT model is suggested as an additional tool for the examination of item-level regional DIF in the context of ILSAs or other large-scale assessments where national and subnational comparisons are intended. Compared to other multiple group DIF detection methods, the distinctive feature of this localized IRT approach to regional DIF detection is the consideration of and accounting for the underlying spatial structure of observations.

The purpose of the current study is to describe and to illustrate the proposed method by providing several detailed examples in the form of simulated case studies and one empirical application. Besides detailing modeling choices such as the use of the 1PL or 2PL model and the selection of a bandwidth to smooth the surface of the parameter estimates, manipulated factors in the case studies include the magnitude and nature of DIF, the spatial arrangement of groups exhibiting DIF, and the local sample sizes.

The current manuscript first provides a background on local spatial modeling techniques including Geographically Weighted Regression (GWR). In addition, Item Response Theory (IRT) modeling techniques are discussed including the model types, model parameters, model parameter estimation techniques, and model assumptions. Then, background on Differential Item Functioning (DIF) is provided including working definitions of both uniform and non-uniform DIF in the context of this paper and a discussion of both traditional DIF detection techniques and new DIF detection techniques extended to the multi-group setting. This literature review will not be a comprehensive overview of local spatial modeling, item response theory, or DIF detection
techniques but is intended to give the reader the necessary theoretical background and motivation for the proposed localized IRT modeling method of regional DIF detection.
Chapter 2

Literature Review

The purpose of the current study is to both propose a localized approach to IRT modeling that could detect item-level regional DIF, while accounting for spatial structure, and to illustrate the proposed methodology using examples that will demonstrate the application of the method and guide procedural choices made by a practitioner. To provide the necessary, though not exhaustive, background and motivation for the currently proposed regional DIF detection method, background on local spatial modeling, IRT modeling, and DIF detection techniques is provided.

Overview of Spatial Data Analysis

As defined in Cressie, spatial data are distinct from other data forms in that they are a realization of a spatial stochastic process \( \{ Y(s): s \in \mathcal{D} \} \) where \( s \in \mathbb{R}^d \) represents the location where data are observed and \( \mathcal{D} \) is a random set in \( d \)-dimensional Euclidean space (1993, p.8). Spatial data sets can be further classified as either point-referenced data, areal data, or point pattern data (Banjeree, Carlin & Gelfand, 2015, p. 2). These three classifications of spatial data sets are also known as geostatistical data, lattice data, and spatial point pattern data (Cressie, 1993, p. 8-9). Statistical modeling approaches differ depending upon the spatial data classification. However, as the focus of current work will be on areal or lattice data, this will be the only one of the three spatial data set classifications formally defined. Lattice (or areal) data consist of measurements in \( \mathcal{D} \) where \( \mathcal{D} \) is a fixed subset of \( \mathbb{R}^d \) (of regular or irregular shape) partitioned into a finite or countably infinite number of areal units with well-defined boundaries (Cressie, 1993, p. 8; Banjeree et al., 2015, p. 2). When observed spatial data occur at locations equally spaced throughout the region \( \mathcal{D} \) this is referred to as regular lattice data. When observed
spatial data occur at locations unequally spaced throughout the region $\mathcal{D}$ this is referred to as irregular lattice data. For instance, irregular lattice data may consist of measurements aggregated at the county level such as for the 159 counties in Georgia. Observed measurements might be associated with geographic locations set at either the county seat or at the county centroid, as in Figure 3. However, irregular lattice data may consist of observations aggregated on smaller areal units such as the neighborhood or district level, as seen in Figure 4 where the 77 neighborhoods/districts of Chicago, Illinois are presented.

Figure 3. Locations of Georgia county centroids and county seats.
Educational data, in the context of ILSAs, is typically aggregated into blocks representing schools, school districts, states/territories, or countries, such that the lattice data examples provided in Figure 3 and Figure 4 are commonplace. Spatial association is introduced into this lattice data by defining a neighborhood structure represented by a proximity or contiguity matrix (Banjeree et al., 2015, p. 74). The ‘neighborhood structure’ can be defined in a variety of ways for both regular and irregular lattice data. When working with regular lattice data, it is common to define the neighborhood structure based upon shared boundaries (e.g., shared border edges or vertices). Due to the equally spaced, grid-like nature of regular lattice data, common neighborhood structures are defined using chess-like language. For example, a neighborhood
structure can be defined using the Rook’s, Bishop’s, or Queen’s case respectively (see Figure 5). To form a proximity or contiguity matrix, those areal units that are within the neighborhood of a specific spatial site could be given a unit weight whereas those areal units that are not within the neighborhood would be given a zero weight. These weights could also reflect, in some form, the “distance” between areal units where distance might be defined using any distance metric. Typically, this distance metric is Euclidean distance though it need not be and could be a Minkowski distance metric or a great-circle/geodesic distance metric, which accounts for the curvature of Earth. The neighborhood structure for irregular lattice data can also be formed using ideas of shared borders or similar to a K-nearest neighbors framework. However, typically in irregular lattice data, neighborhood structures are based upon distance measurements from areal unit centroids or other areal unit locations (e.g., county seats). For example, Figure 6 demonstrates the neighborhood structure that is created for Georgia counties when the K-nearest neighbors framework is implemented ($k = 4$) and when the nearest neighbors within a certain

![Figure 5. Neighborhood structure possibilities for regular lattice data.](image-url)
Euclidian distance, D-nearest neighbors framework, is implemented ($d = 50$ km). Moreover, the idea of utilizing a D-nearest neighbors framework in creating a neighborhood structure for irregular lattice data can be seen in Figure 7 where spatial ‘neighborhoods’ are created for the Chicago, Illinois neighborhoods/districts using a specified distance metric as the radius of a circular neighborhood structure. The spatial ‘neighborhood’ is denoted by areal units that fall within the solid green circle with a radius defined by a distance metric denoted by a blue dotted line. For visual clarity in illustrating the construction of spatial ‘neighborhoods’, only seven of the more than seventy spatial ‘neighborhoods’ are drawn. The idea of creating neighborhoods based on a distance metric is especially important when spatial data sets involve a mixture of both point-referenced and lattice (areal) data. For instance, in Figure 8 one may consider both the spatial locations of measurements taken at the red points (representing housing prices at certain locations) and the spatial locations of measurements taken at an aggregate borough-level in the city of London, United Kingdom (Lu et al., 2017).

Figure 6. Neighborhood structure possibilities for Georgia counties.
Figure 7. Neighborhood structure possibilities for Chicago neighborhoods/districts.

Figure 8. Mixture of point-referenced and irregular lattice spatial data for the London boroughs.
As discussed previously, weights in a proximity or contiguity matrix could be binary whereby those areal units that are within the neighborhood of a specific spatial site are given a unit weight and those areal units that are not within the neighborhood are given a zero weight. However, weights in a proximity or contiguity matrix can reflect the distance that exists between two areal units within a neighborhood structure such that nearer areal units are assigned greater weights and more distant areal units are assigned lesser weights. Referring to Figure 9, the assigned weights could decay to a zero weight as a function of distance from the spatial site or location of interest. Consequently, neighboring areal units in close proximity to the spatial site have higher weights that lessen as the distance between areal units and the spatial site, represented by the widening green circles in Figure 9, increases. The assignment of weights produce neighborhood structures that can be referred to as discrete or fuzzy zones illustrated in Figure 10. Weights that are either binary or decay as a discontinuous function of distance, so that units beyond a specified point are assigned a zero weight, create discrete zoning systems for

\[ X = \text{Spatial Site/Location or} \]

“Regression” Point

*Figure 9. Proximity/contiguity matrix weights.*


neighborhood structures. Alternatively, weights that decay as a continuous function of distance, such as assigning weights based upon a Gaussian or exponential weighting scheme, create fuzzy zoning systems for neighborhood structures.

![Discrete vs Fuzzy Zoning Systems](image)

*Figure 10.* Illustration of discrete and fuzzy zoning systems.

Once a neighborhood structure is defined for lattice spatial data, models can be considered that incorporate the spatial structure, as established by the defined spatial neighborhood and the proximity/contiguity matrix (Banjeree et al., 2015). Two popular global spatial models are the simultaneous autoregressive (SAR) model developed by Whittle and the conditionally autoregressive (CAR) model developed by Besag (as cited in Banjeree et al., 2015, p. 5). These global models will not be thoroughly discussed as the current work intends to propose a local modeling approach.

Spatial data are not presumed to be independent, with measurements that are closer together in space sometimes being more related to one another than observations at a distance
(Tobler, 1970). In fact, observations close together in space might be more or less related to one another than observations at a distance. This suggests a spatial correlational structure that needs to be incorporated into models. A similar concept is utilized in time series analysis where models account for the autocorrelation that exists between observations taken in time. Spatial autocorrelation is a measure of the correlation of a variable with itself throughout space. Positive spatial autocorrelation suggests that observations are more similar when near to one another whereas negative spatial autocorrelation suggests that observations are more dissimilar when near to one another. The strength of the spatial autocorrelation among areal units on a global level can be formally measured by statistics such as Moran’s I and Geary’s C, analogues to measures of association occurring in time series analysis (Banjeree et al., 2015, p. 75).

According to Brunsdon and Comber (2015), the more common measure of the two is Moran’s I and, consequently, this measure is detailed below. Moran’s I can be formulated as in Equation 1 (Banjeree et al., 2015, p. 75):

\[
I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i \neq j} w_{ij} \sum_i (Y_i - \bar{Y})^2}
\]

(1)

where \(w_{ij}\) is the spatial weight feature between areal units \(i\) and \(j\). However, as noted by Brunsdon and Comber (2015), \(w_{ij}\) can be a binary indicator of whether areal units \(i\) and \(j\) are neighbors, taking the value of 0 if they are not neighbors and the value of \(\frac{1}{|\delta_i|}\) with \(|\delta_i|\) being the number of areal unit neighbors for areal unit \(i\). Moran’s I is a correlation of a variable with itself in a sense, as it is the correlation of a variable with the spatial lag of that variable found by averaging over all neighboring areal units. Moran’s I, as such, is similar to Pearson’s linear correlation coefficient in many respects but is not bounded on \([-1,+1]\) due to the incorporation of spatial weights. Despite this difference, interpretation of Moran’s I is similar to that of Pearson’s
linear correlation coefficient but in a spatial data context. If Moran’s I is a positive value, this indicates that neighboring areal units have similarly low or high values of a measured variable. As such, positive values would indicate spatial clustering of areal units. If Moran’s I is a negative value, this indicates that neighboring areal units have dissimilarly low or high values of a measured variable. As such, negative values would indicate spatial dispersion of areal units. If Moran’s I is a value very close to zero, this indicates that there is no spatial autocorrelation/association present among the areal units for the measured variable. Significance testing under the null hypothesis of no spatial autocorrelation can be performed with a test statistic utilizing an approximate normal distribution or with a permutation test (Brunsdon & Comber, 2015). Moreover, Anselin (1995) suggested that the spatial relationship summarized by Moran’s I can be visualized by utilizing a lagged mean plot, also known as a Moran plot or Moran scatterplot. In a Moran scatterplot, the value for a measured variable in each areal unit is compared to the weighted average of the measured variable values for neighboring areal units. In fact, Moran’s I is the slope of a linear regression of the lagged means (i.e., the weighted averages of values for neighbors) against the mean values for each areal unit in a spatial region. As such, any observed outliers in a Moran scatterplot might be functioning as leverage points thereby indicating local spatial patterns in the data that might be unobservable at an aggregate level.

Spatial data analysis is distinct from other forms of statistical data analysis due to the correspondence of observations with some fixed or random location in geographic space that can be represented by a set of coordinates, such as longitude and latitude. While sharing similarities with time series data analytical techniques, such as accounting for autocorrelation in observations, spatial autocorrelation is slightly more complex owing to the difference in the concept of a time lag (past, present, future steps) compared to that of a spatial lag. Different
types of spatial data classifications, neighborhood structure definitions, and weighting schemes based upon various distance metrics are taken into consideration in spatial models such as the global SAR and CAR models mentioned previously. With a foundational understanding of global spatial data analysis, the concept of local spatial data analysis can now be discussed.

**Local Spatial Data Analysis**

While local approaches to data analysis are not new, the further development and use of these local modeling techniques as well as the application of such local modeling techniques to spatial data analysis across a variety of disciplines has steadily risen making use of recent advances in geographic information systems (GIS) and recent increases in geographic data collection (Fotheringham, Brunsdon & Charlton, 2002). While some of these local spatial modeling techniques will be discussed in this manuscript, a more thorough overview of local spatial techniques may be found in Lloyd (as cited in Matthews & Yang, 2012).

Local spatial models differ from global spatial models in a variety of ways, as will be discussed below. However, it is important to define what is meant by the terms ‘global’ and ‘local’. Global spatial models are statements about spatial processes which are assumed to be stationary over the study region and, as such, are location independent. Local spatial models, on the other hand, are spatial decompositions of global models that focus on subsets of data and that allow spatial processes to exhibit nonstationarity over the study region; as such, local spatial models are location dependent.

With global and local spatial models now defined, the reasons for utilizing local spatial models, the early development of local spatial models, the descriptions of three common local spatial models and the most frequently used local indicator of spatial association (LISA) will be discussed.
**Reason for local spatial models.** Global models such as generalized linear models (GLM) necessarily assume that the relationship between explanatory and response variables is homogeneous or stationary across a spatial region. However, GLM models do not account for spatial autocorrelation and are not typically used when analyzing spatial data. To account for spatial autocorrelation in data, two spatial modeling techniques (SAR and CAR models) are commonly applied. However, SAR and CAR models are global models. While these models, with proper specification, can account for spatial autocorrelation in the variables in the model and in the model residuals, the relationships being modeled are still assumed to be the same everywhere across a spatial region, depending only on a spatial lag rather than a specific spatial location. Global models assume spatial homogeneity or stationarity.

The assumption of spatial homogeneity or stationarity in global models may be violated in practice however. For instance, violations of spatial stationarity (i.e., spatial nonstationarity) might result from sampling variation, model misspecification, or the existence of relationships that intrinsically differ across a spatial region (Fotheringham et al., 2002, p. 9-10). If spatial stationarity is violated, the use of global models is inappropriate and may not accurately reflect the relationships between variables that are being modeled. Consider the ecological fallacy whereby inferences about individuals and relationships between variables on an individual level are made based upon observed relationships between these same variables on an aggregate level (Banjeree et al., 2015, p. 165). This may result in the spatial equivalent of Simpson’s paradox, which refers to the reversal of inferential findings when data is analyzed in aggregate as opposed to in a disaggregate form (as cited in Fotheringham et al., 2002, p.7-8). Global models, as such, might obscure or hide spatial differences in variable relationships whereas Fotheringham et al.
likened local models to ‘spatial microscopes’ that can uncover these previously hidden spatial patterns (as cited in Matthews & Yang, 2012).

A related issue that is discussed thoroughly in a review paper by Gotway and Young, is the modifiable areal unit problem (MAUP) where results of global analysis are seen to depend upon the level of spatial aggregation upon which data are collected (as cited in Banjeree et al., 2015, p. 165). Also known as a zone definition problem, there are two separate components of the MAUP: the scale effect and the zoning effect (Fotheringham et al., 2002, p. 144). The scale effect refers to the idea that the same statistical analysis can produce differing and, at times, conflicting results when the models are calibrated at different spatial resolution levels (Fotheringham et al., 2002, p. 144). The zoning effect refers to the idea that statistical analyses can produce differing and, at times, conflicting results when the different statistical models are calibrated to different groupings of zones at the same spatial resolution level (Fotheringham et al., 2002, p. 144). Local models allow for an analysis of the sensitivity and/or stability of spatial model parameter estimates by allowing model refitting over a wide range of data aggregation levels and zoning systems as well as the visualization of estimated parameter sensitivity and/or stability. Global models, however, do not inherently attempt to address the MAUP in that they assume spatial stationarity and, although local models cannot eliminate the MAUP issues, Fotheringham et al. (2002, p. 153) argue that local models may be less influenced by these scale and zoning effects than their global model counterparts.

Besides modeling spatial nonstationarity directly and minimizing the influence of MAUP issues, local models that disaggregate spatial data allow links to GIS. By producing parameter estimates that are location dependent, results of local models are mappable providing immediate visualization of patterns in a spatial region (Fotheringham et al., 2002, p. 25). Visualizations of
underlying local spatial patterns, which are not possible for global models, can facilitate interpretation and can suggest subsequent analyses (Matthews & Yang, 2012).

**Local indicators of spatial association.** Local models are spatial decompositions of global models and, as such, can reveal spatial nonstationarity. Spatial autocorrelation has been discussed previously as well as measures of spatial autocorrelation such as Moran’s I (see Equation 1). However, as previously defined, Moran’s I is a measure of global spatial autocorrelation, summarizing the extent to which observed values are more similar or more dissimilar on average when near to one another in space. In order to measure the extent to which observed values are more similar or more dissimilar when close to a specific location in space, a decomposition of Moran’s I is necessary. Local indicators of spatial association (LISAs) as defined by Anselin (1995) include statistics such as Local Moran’s I and Local Geary’s C. These local decompositions of the global spatial autocorrelation statistics provide a way to identify local effects (e.g., clusters and hotspots) and spatial nonstationarity. Local Moran’s I can be formulated as in Equation 2 (Anselin, 1995):

\[ I_i = z_i \sum w_{ij} z_j \]  

(2)

where \( w_{ij} \) is the spatial weight feature between areal units \( i \) and \( j \) and \( z_i, z_j \) are the mean centered values of the original variable (i.e., the deviations of the original variable for areal unit \( i \) and \( j \) from the mean, \( Y_i - \bar{Y} \) and \( Y_j - \bar{Y} \) respectively).

Local Moran’s I can also be formulated as in Equation 3 where the previous form (see Equation 2) is divided by the sample variance for all areal units \( k = 1, ..., n \) of the original variable (Bivand, 2017):

\[ I_i = \frac{(Y_i - \bar{Y}) \sum_{j=1}^{n} w_{ij}(Y_j - \bar{Y})}{\sum_{k=1}^{n}(Y_k - \bar{Y})^2/(n-1)} \]  

(3)
With either of the above formulations, Equation 2 or Equation 3, the Local Moran’s I value for each location in space indicates if there is significant clustering of similar or dissimilar values around a particular point with significance testing possible (Anselin, 1995). Consequently, the LISA value serves to detect both spatial clusters and spatial hotspots. If the Local Moran’s I value for a particular point is a positive value, this indicates that neighboring areal units have similarly low or high values of a measured variable. As such, positive Local Moran’s I values would indicate spatial clustering of areal units around a particular location, \( i \). If the Local Moran’s I value for a particular point is a negative value, this indicates that neighboring areal units have dissimilarly low or high values of a measured variable. As such, negative Local Moran’s I values would indicate spatial dispersion of areal units around a particular location, \( i \), which would appear to be a ‘hotspot’ or ‘outlier’. If the Local Moran’s I value for a particular point is very close to zero, this indicates that there is no local spatial autocorrelation/association present at a specified spatial location for the measured variable.

Since LISAs provide several statistics, one for each areal unit in a spatial region, LISA values can be mapped to reveal stronger or lesser local spatial autocorrelation. Moreover, as mentioned above, significance testing is possible but Bonferroni adjustment or some other type of multiple testing adjustment is necessary.

According to Anselin (1995), LISAs uncover hidden local spatial patterns that global statistics average over, avoiding ecological fallacy. While a global statistic at a given spatial lag may be statistically significant, without the calculation of a LISA such as Local Moran’s I, large areas of no spatial autocorrelation and the existence of certain locations with large leverage may be hidden. Moreover, while a global statistic may be statistically insignificant, LISA values might reveal hidden areas of local spatial autocorrelation.
For good reason, there is an ever-increasing exploration of potential hidden local patterns with the use of local spatial modeling techniques. Some of these local spatial modeling techniques are discussed below, some in brief and some in detail.

**Early local spatial models.** Finley (2011) notes that Geographically Weighted Regression (GWR) techniques and Spatially Varying Coefficients (SVC) hierarchical modeling techniques, the latter of which are often specified in a Bayesian framework (B-SVC), are currently the most often used methods for modeling data that exhibits spatial nonstationarity. However, numerous methods for addressing spatial nonstationarity have previously been proposed, inspired in large part by the Random Coefficient Model (RCM) described by Rao (as cited in Charlton & Fotheringham, 2009) and by the Varying Coefficient Model (VCM) described in Hastie and Tibshirani (1993). Owing to the fact that GWR is a special case of the Hastie and Tibshirani’s VCM (Fotheringham et al., 2002, p. 87), this broad class of models as well as several of the previously proposed methodologies to address spatial nonstationarity will be discussed in brief.

**Varying coefficient models.** As described by Hastie and Tibshirani (1993), a Varying Coefficient Model (VCM) represents a broad class of models, defining a framework that encompasses such models as GLM, Generalized Additive Models (GAM), and Dynamic Generalized Linear Models (Generalized DLM). The presentation of this broad class of models, which “allow the coefficients that describe the effect of a regressor to vary as a function of another factor,” extended generalized regression modeling techniques (Hastie & Tibshirani, 1993, p. 774). Moreover, VCM provided a framework to model spatial nonstationarity that would serve as inspiration for subsequently proposed local spatial modeling techniques, though this was not originally suggested by the authors.
Hastie and Tibshirani (1993, p. 757) suppose that for a random variable $Y$ whose distribution depends on a parameter $\eta$, two types of predictors, $X_i$ and $R_i$, and unspecified functions $\beta_i()$ with $i = 1, \ldots, p$, a VCM has the following form:

$$\eta = \beta_0 + X_1\beta_1(R_1) + \ldots + X_p\beta_p(R_p)$$  \hspace{1cm} (4)

The dependence of the unspecified functions $\beta_i()$ on $R_i$ implies an interaction between the predictors $X_i$ and $R_i$. With very little restriction on the unspecified functions in a VCM model, Equation 4 specifies a broad framework encompassing many modeling techniques. For instance, Hastie and Tibshirani (1993, p. 761) note that if $\beta_i(R_i) = \beta_i$ for $i = 1, \ldots, p$, then each term is linear in $X_i$ and Equation 4 would represent a GLM. Also, if $X_i = c$ then each term in the VCM model is simply an unspecified function in $R_i$ making Equation 4 represent a GAM (Hastie & Tibshirani, p. 761). Moreover, if the $R_i$s are the same variable, a factor such as time, which modifies the effects of $X_i$ for $i = 1, \ldots, p$, then Equation 4 could be modeled as a Generalized DLM (Hastie & Tibshirani, p. 762).

The VCM as specified by Hastie & Tibshirani (1993) has broad applications and provides an overarching framework connecting many models including, but not limited to, the GLM, GAM, and Generalized DLM models described above. This broad class of VCM models allows the relationships between explanatory and response variables to vary as a function of another factor and, as such, implies an interaction between predictors $X_i$ and $R_i$. Essentially, Hastie & Tibshirani (1993) detail a family of models that allows for the study of interactions, for the study of nonstationary relationships, necessarily making the VCM framework one which is appropriate to model spatial nonstationarity.

Other approaches to spatial nonstationarity. To mention only a few areas of application, spatial nonstationarity arises in analyses related to health care delivery, infectious disease,
environmental equity and conditions, housing markets, industrialization and development, population density, poverty, religion, traffic, and voting (Matthews & Yang, 2012). To match the growing need for models that would address spatial nonstationarity, several models have been proposed since the 1970s. For instance, Swamy proposed an extension of Rao’s 1965 RCM to the spatial case in 1971 (as cited in Charlton & Fotheringham, 2009). In order to address spatial nonstationarity and allow parameters to vary over a geographic space, the Expansion Method of Casetti was extended to the spatial case by several researchers (Fotheringham et al., 2002, p. 16). Foster and Gorr proposed Spatial Adaptive Filtering in 1986 (as cited in Charlton & Fotheringham, 2009).

Spatial nonstationarity can in many ways be addressed through the use of multi-level modeling as suggested by Goldstein in 1987 (as cited in Charlton & Fotheringham, 2009). This methodology is not unfamiliar in educational research. Multi-level modeling attempts to avoid both the ecological fallacy discussed previously as well as the atomistic fallacy, wherein behavior observed at only a disaggregate level is missing the context in which this behavior occurs, by combining a hierarchy of at least two levels that might represent behavior at a disaggregate level and, also, behavior at an aggregate level. Though multi-level modeling is both familiar in educational research and also seems appropriate for addressing spatial nonstationarity, this type of modeling technique relies on a priori definitions of spatial units at each level of the hierarchy. Consequently, multi-level modeling assumes the spatial process is discontinuous at pre-defined spatial boundaries and does not utilize underlying spatial structure (Fotheringham et al., 2002, p. 19).

SVC models are related to VCM, multi-level modeling, and hierarchical modeling techniques. They are often specified in a Bayesian framework (commonly referred to as
Bayesian Spatially Varying Coefficient models, B-SVC). B-SVC models are one of the most common current methods to address spatial nonstationarity (Banerjee et al., 2015; Lloyd, 2010) however these models will not serve a role in the context of the current research and, for that reason, will not be discussed.

Geographically Weighted Regression

Geographically weighted (GW) models comprise a broad class of spatial modeling techniques that attempt to address spatial nonstationarity through the local calibration of estimated model parameters (Gollini, et al., 2015). Geographically weighted regression (GWR) is one of the many GW models currently used for local spatial modeling across a variety of disciplines. GWR is an exploratory technique which generates a set of location-specific parameter estimates by utilizing a moving window approach originally inspired by the idea of LOESS smoothing as seen in Cleaveland (1979). As previously discussed, GWR is also a special case of Hastie and Tibshirani’s VCM (Fotheringham et al., 2002, p. 87). Motivating the idea of a localized approach to IRT modeling, GWR will be described below. The precursor to GWR, moving window regression, will be discussed as will the model specifications for GWR, the spatial weighting function for GWR, the extensions of GWR, and the issues inherent in GWR modeling. The current discussion of GWR modeling will be brief though the methodology is presented in Lloyd (2010) and fully described in Fotheringham et al. (2002).

Precursor to geographically weighted regression. Prior to the full development and presentation of GW modeling techniques, including GWR, attempts to model spatial nonstationarity were limited to areal unit calibrations of global models. In this way, the global models could be decomposed by calibrating the global model itself separately to each of the smaller, more localized areal units in the spatial region of interest (Fotheringham et al., 2002, p.
38). However, this separate calibration of the global model on each of the areal units in a spatial region assumes a discontinuous spatial process that does not take into consideration the similarities of neighboring areal units or of areal units that lie in close proximity to one another. This separate calibration technique also necessarily assumes the importance of the administrative and political boundaries which typically define the areal units such as the boundaries of countries, states, or territories. Moreover, the idea of separate calibrations of global models for pre-specified areal units without accounting for the proximity of said areal units calls to mind the MAUP issues previously discussed.

To address some of the abovementioned issues, moving window regression does not rely on pre-defined areal units but, rather, utilizes regions that are often square or circular in shape to sweep across a spatial region of interest. Global models are calibrated several times for each of these regions, which are centered at several ‘fit’ or ‘regression’ points that need not be specified at locations where data were collected. Moving window regression essentially involves the repeated calibration of several models, each using only a subset of observations that lie within these square or circular spatial regions. While this does not entirely solve MAUP issues as the technique still models a discontinuous spatial process, this moving window approach incorporates some level of spatial structure, allows for spatial changes in the estimated parameters to be monitored and, by modifying the area of the moving windows (i.e., the defined spatial regions), results in a smoother surface of parameter estimates (Fotheringham et al., 2002, p. 42-43). Moving window regression utilizing circular spatial regions, which have either a fixed or an adaptive radius, is actually a special case of GWR making use of a discrete zoning system and a binary weighting scheme.
**Model.** GWR utilizes a moving window approach to localized spatial modeling. Regression points are first specified for a spatial region, often by overlaying a grid upon an irregular lattice and choosing several equidistant locations for model calibration. An example of how one might select regression points even with a mixture of point-referenced and irregular lattice spatial data can be seen in Figure 11. In the case of lattice spatial data, regression points are often allocated to areal unit centroids. Similar to moving window regression, all observations that lie within a certain distance of a specified regression point are included in model calibration and the process is repeated for each regression point in a spatial region of interest (Fotheringham et al., 2002, p. 44). However, unlike moving window regression, the weights of these observations used for model calibration need not be confined to a binary weighting scheme.

*Figure 11.* Overlay grid of the London boroughs for regression point selection.
Weights can be assigned for each model calibration so that they decrease continuously as a function of the distance between an observation and a regression point.

GWR techniques apply the idea of local modeling and moving window techniques to linear regression in order to analyze spatial nonstationarity, which is one assumption of the GWR model. The nonstationarity is directly addressed by allowing the relationships between predictor variables and the outcome variable to change over space. Separate regression models are created at each regression point and model coefficients are estimated utilizing a method similar to weighted least squares (WLS) that applies a spatial weights matrix conditioned on the individual location, \((u_i, v_i)\). Fotheringham et al. (2002, p. 52) suppose that for a dependent variable \(Y\) and a set of \(m\) independent variables, \(X_k\) where \(k = 1, \ldots, m\), each of the \(n\) observations in a dataset have a measurement of spatial position available in a suitable coordinate system so that the \(i\)th point in space has coordinates denoted by \((u_i, v_i)\). Assuming the above, a GWR model has the following form specific to location \(u\), where \(u\) is a vector of the coordinates \((u_i, v_i)\):

\[
Y_i(u) = \beta_{0i}(u) + \beta_{1i}(u)X_{1i} + \ldots + \beta_{mi}(u)X_{mi} + \varepsilon_i
\]

As specified by Fotheringham et al. (2002, p. 53), estimates of \(\beta_k\) coefficients are based upon weights conditioned on the specific location \((u_i, v_i)\) and are calculated as in Equation 6 below:

\[
\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) Y
\]

where the bold type in Equation 6 denotes a matrix. Notice that \(X^T W(u_i, v_i) X\) is the geographically weighted variance-covariance matrix and \(W(u_i, v_i)\) is an \(n\) by \(n\) diagonal spatial weighting matrix of the following form (Equation 7):

\[
W(u_i, v_i) = \begin{bmatrix}
w_{i1} & \ldots & 0 \\
\vdots & \ddots & \vdots \\
0 & \ldots & w_{in}
\end{bmatrix}
\]
Each diagonal element $w_{ij}$ denotes the weight given to data point $j$ for $j = 1, \ldots, n$ in the calibration of the model for regression point $i$ with coordinates $(u_i, v_i)$. Equation 6 represents a WLS estimator but with a weight matrix that varies according to the specific location of regression point $i$ having coordinates $(u_i, v_i)$. Consequently, the weighting matrix is computed for each model calibration and the weights themselves are specified according to a weighting scheme, also known as a spatial weighting function (Fotheringham et al., 2002, p. 53-54).

Weighting schemes can be based upon one of several different distance metrics, can utilize either discontinuous or continuous weighting functions, and can be fixed or adaptive.

**Spatial weighting function.** Gollini et al. (2015) note that the spatial weighting function is a fundamental aspect of GW modeling. The spatial weighting function quantifies the spatial relationship between observed variables and results in a diagonal spatial weighting matrix, $W(u_i, v_i)$, which is location dependent and, as such, is computed for each model calibration at the specified regression points. The diagonal elements, $w_{ij}$, of matrix, $W(u_i, v_i)$, depend upon specification of the distance metric, the kernel function, and the bandwidth (Gollini et al., 2015).

The distinct feature of GWR as compared to moving window regression is that spatial weights can decay as a function of the distance between observed data points and the regression point used for model calibration. Consequently the distance between the $j$th observation and the regression point at location $i$, denoted $d_{ij}$, must be measured. As discussed previously, the “distance” between any two spatial locations may be defined using any distance metric. Typically, $d_{ij}$, the distance between location $i$ and $j$, is measured as Euclidean distance. However, GW models (including GWR) can measure $d_{ij}$ in terms of other distance metrics such as the great-circle/geodesic distance metric, which accounts for the curvature of Earth.
The kernel function dictates the manner in which weights, \( w_{ij} \), are calculated based upon the distance between the two locations, \( d_{ij} \), and a bandwidth, \( b \). Six kernel functions are provided in Table 1. The Global model kernel function is included, as in Gollini (2015), to indicate that the global model is a special case of the GWR local model. In addition, the box-car kernel represents moving window regression and is included to indicate that moving window regression is a special case of the GWR local model.

Table 1. *Kernel functions available for GW modeling.*

<table>
<thead>
<tr>
<th>Continuous Kernel Functions</th>
<th>Global Model ( w_{ij} = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian ( w_{ij} = \exp \left( -\frac{1}{2} \left( \frac{d_{ij}}{b} \right)^2 \right) )</td>
<td></td>
</tr>
<tr>
<td>Exponential ( w_{ij} = \exp \left( -\sqrt[n]{d_{ij}} \right) )</td>
<td></td>
</tr>
</tbody>
</table>

| Discontinuous Kernel Functions | Box-car \( w_{ij} = \begin{cases} 1 & \text{if } |d_{ij}| < b \\ 0 & \text{otherwise} \end{cases} \) |
|--------------------------------|-------------------------------------------------------------|
| Bi-square \( w_{ij} = \begin{cases} \left( 1 - \left( \frac{d_{ij}}{b} \right)^2 \right)^2 & \text{if } |d_{ij}| < b \\ 0 & \text{otherwise} \end{cases} \) |
| Tri-cube \( w_{ij} = \begin{cases} \left( 1 - \left( \frac{|d_{ij}|}{b} \right)^3 \right)^3 & \text{if } |d_{ij}| < b \\ 0 & \text{otherwise} \end{cases} \) |
Kernel functions for GW modeling can be either continuous or discontinuous functions. As seen in Table 1, the Gaussian and exponential kernel functions are continuous functions that provide the maximum weight for an observation at the location \( j \) when location \( j \) corresponds to location \( i \), the regression point. The weights then decay continuously as a function of the increasing distance between \( i \) and \( j \). These continuous kernel functions produce fuzzing zoning systems for spatial neighborhood structures. In fact, Fotheringham et al. (2002, p. 129) state that GWR can be “seen as a technique for allowing fuzzy zones to be placed around each regression point.” However, there are also three commonly used kernel functions provided in Table 1 that are discontinuous, producing discrete zoning systems for spatial neighborhood structures whereby the weights \( w_{ij} \) are set to be zero when the distance between \( i \) and \( j \), \( d_{ij} \), exceeds the bandwidth, \( b \). The bi-square and tri-cube kernel functions, though discontinuous, are still distance-decay weighting kernels.

Unlike the other specified kernel functions, the box-car kernel function is not a distance-decay weighting kernel. The box-car kernel corresponds to moving window regression and necessarily assumes a discontinuous spatial process. Despite these limitations, Gollini et al. (2015) note that the box-car kernel function is more computationally efficient than distance-decay weighting kernels and therefore is more useful when dealing with large datasets. Moreover, Fotheringham et al. (2002, p. 44) mention that “the results of GWR are relatively insensitive to the choice of weighting function but they are sensitive to the bandwidth”. In general, the shape of the kernel function has less influence on the GW model fitting than does the choice of the bandwidth (Charlton & Fotheringham, 2009).

Gollini et al. (2015) point out that the bandwidth is the key controlling parameter in all of the kernel functions in Table 1. The bandwidth, \( b \), can be thought of as a smoothing parameter
with larger bandwidths corresponding to greater smoothing of the parameter estimate surface (Fotheringham et al., 2002, p. 211). The bandwidth can be specified as a fixed distance or as an “adaptive” distance, which would provide a fixed number of local data observations so that local sample size can be fixed for each model calibration (Gollini et al., 2015). Adaptive bandwidths are utilized when sample data points are not regularly spaced throughout the spatial area of interest. Due to the non-regularity of the observed data points, some local regressions are calibrated on a very limited number of data points resulting in large standard errors or in a failure for parameter estimation around a specific regression point, leading to an undersmoothed surface of parameter estimates (Fotheringham et al., 2002, p. 46). In these instances, an adaptive bandwidth can be utilized to accommodate for the clustering and irregularity in the number of data points around a given regression point location. If the observed data points are approximately regularly spaced, however, a fixed bandwidth is reasonable and suitable for GW modeling. While optimal bandwidth selection methods have been proposed for various GW models such as minimizing the “corrected” AIC or the Cross-Validation (CV) score, Fotheringham et al. (2002) argue for the utilization of several bandwidths. Rather than relying on one pre-specified bandwidth, utilizing several bandwidths provides for model calibration across a wide range of data aggregation levels and allows for the sensitivity of spatial model parameter estimates to be investigated thereby minimizing the impact of MAUP issues (Fotheringham et al., 2002, p. 153).

GW models have now extended to include modeling techniques such as but not limited to GW generalized linear models (GWGLM), GW principal component analysis (GW PCA), GW ridge regression, mixed GWR, heteroskedastic GWR, and GW discriminant analysis (Gollini et al., 2015). Despite concerns with GW modeling techniques such as those related to collinearity,
MAUP, and inferential issues, GW models explore spatially varying relationships on a local level and allow for the visualization of these spatially varying relationships. GW models are powerful exploratory tools with appealing visualization potential that are growing in influence in several disciplines and even promoted in disciplines such as health policy where spatial nonstationarity is suspected (Matthews & Yang, 2012; Mennis, 2006). Consequently, it will be of interest to develop an IRT modeling technique inspired by GW models allowing for the investigation of spatially varying item functionality of survey instruments and educational assessments, especially in the context of ILSAs.

**Item Response Theory**

Item response theory (IRT) models are a specific subclass of latent variable models that attempt to link an observed response variable, which may be dichotomous or polytomous, with a latent trait that represents an unobserved variable describing the extent to which an individual possesses a certain property such as mathematical ability, brand adherence, political support, or anxiety (Rizopoulos, 2006). While many applications of IRT are found in educational assessment, this modeling framework can be applied to a wider class of measurement problems in a variety of fields ranging from education to psychology, sociology, marketing, political science, and public health (Fischer & Molenaar, 1995, p. 3; Rizopoulos, 2006). Noting both the increasing importance of ILSAs in educational research for making cross-national or subnational comparisons of a measured latent trait, namely examinee ability level, and also the possibility of DIF due to geographic or spatial location, IRT modeling will be described below to provide the necessary background for proposing a localized approach to such models.

As previously discussed, IRT is a modeling technique belonging to the broader class of latent variable models (Rizopoulos, 2006). The general form for latent variable models will be
discussed as will the main IRT model types, the model parameters, the primary model parameter estimation methods, and the underlying model assumptions.

**Latent variable models.** As described in Rizopoulous (2006), a latent variable regression model may have the following form (Equation 8) assuming a given set of response variables, \( x_1 \ldots x_p \), and a set of latent variables, \( z_1 \ldots z_q \), (where \( q \ll p \)):

\[
E(x_i | z) = g(\lambda_{i0} + \lambda_{i1}z_1 + \ldots + \lambda_{iq}z_q)
\]

(8)

where \( g(\cdot) \) is a link function, \( \lambda_{i0} \ldots \lambda_{iq} \) are regression coefficients for the \( i \)th manifest variable, and the response variables are conditionally independent given the latent variables. If one considers normally distributed continuous response variables with \( g(\cdot) \) being the identity link function in Equation 8, the common factor analysis (FA) model can be seen. If one considers dichotomous or polytomous response variables with one latent variable assumed (although more can be considered in practice) and consider \( E(x_i | z) \) to express the probability of endorsing a particular response category given the latent trait, the basic form for an IRT model emerges (Rizopoulous, 2006). The unidimensional IRT modeling framework for dichotomous data will now be discussed.

**Model types.** Item response theory models for dichotomous data provide a model for the probability of a “correct” response on each of \( k \) items given an assumed latent ability level, \( \theta \). As noted by Fischer and Molenaar (1995, p. 3-4), it is convenient to utilize terminology such as examinees or persons, items, and responses with dichotomous responses scored as “correct” or “wrong” though IRT models can be applied in a wide array of settings other than educational testing. Due to the wide application of IRT, terms such as “examinees,” “items,” and “responses” might refer to different objects in other settings though the IRT model will remain the same. Fischer and Molenaar (1995, p. 4) also note that items scored dichotomously as “correct” or
“wrong” in an educational context can refer to any dichotomous scoring of an item as an affirmative response (correct, positive, agree, high position on the latent trait) or as a non-affirmative response (wrong, negative, disagree, low position on the latent trait).

Rizopoulous (2006) and Hambleton and Swaminathan (1985, p. 37-38) both provide a general model for the probability of a correct response on the $i$th item for the $m$th examinee with a person-specific ability level, $\theta_m$, that has the following form (Equation 9):

$$P(x_{im} = 1|\theta_m) = c_i + (1 - c_i)g\{a_i(\theta_m - b_i)\}$$

(9)

where $x_{im}$ is the dichotomous response on the $i$th item for the $m$th examinee with a corresponding latent ability or skill level of $\theta_m$. Here $g\{\cdot\}$ is a link function, typically a probit or logit link, which corresponds to the normal and logistic metrics of the IRT model. Equation 9 with a logit link is the common software implementation of IRT models. However, a scaling factor of $D = 1.702$ may be used to equate, approximately, the normal and logistic metrics when a logit link is used i.e., $Da_i(\theta_m - b_i)$. For Equation 9, values of $a$, $b$, and $c_i$ represent the discrimination, difficulty, and “guessing” parameters for the $i$th item respectively. The “guessing” parameter typically results in values that are smaller than would be assumed with random guessing and, consequently, is commonly referred to as the pseudo-chance or pseudo-guessing parameter (Hambleton & Swaminathan, 1985, p. 38). These parameters will be described in more detail in the following sections.

Three-parameter logistic model. By incorporating the scaling factor, $D = 1.702$, into Equation 9 and assuming a logit link function, the resulting general form for the unidimensional IRT model may be reformulated as in Equation 10, which corresponds to Hambleton and Swaminathan’s formulation of the three-parameter logistic (3PL) IRT model (1985, p. 37-38).

$$P(x_{im} = 1|\theta_m) = c_i + (1 - c_i)\frac{\exp[Da_i(\theta_m - b_i)]}{1+\exp[Da_i(\theta_m - b_i)]}$$

(10)
Two-parameter logistic model. The 3PL model is numerically less stable than simpler IRT models and de Ayala (2009, p. 131) recommends a calibration sample size exceeding 1000 examinees to mitigate some of the convergence issues that commonly arise in the 3PL model setting. Consequently, the reduction of Equation 10 to the two-parameter logistic (2PL) model provides a reasonable alternative for IRT modeling that is somewhat less flexible but far more stable than the 3PL model. The 2PL model assumes that there is no pseudo-guessing parameter i.e., $c_i = 0$. The resulting IRT model is provided in Equation 11 and corresponds to Hambleton and Swaminathan’s formulation of the 2PL IRT model (1985, p. 36).

$$P(x_{im} = 1|\theta_m) = \frac{\exp[Da_i(\theta_m-b_i)]}{1+\exp[Da_i(\theta_m-b_i)]}$$ (11)

One-parameter logistic and Rasch models. The 3PL model provided in Equation 10 further reduces to the one-parameter logistic (1PL) model when assuming there is no pseudo-guessing parameter and when the discrimination parameter $a_i$ is constant for all $i = 1, ..., k$. The IRT model resulting from these assumptions is provided in Equation 12 and corresponds to Hambleton and Swaminathan’s formulation of the 1PL IRT model (1985, p. 47) where $\bar{a}$ is the common level of discrimination for all items.

$$P(x_{im} = 1|\theta_m) = \frac{\exp[D\bar{a}(\theta_m-b_i)]}{1+\exp[D\bar{a}(\theta_m-b_i)]}$$ (12)

The one-parameter logistic (1PL) model is equivalent to but can be seen as more flexible than the Rasch model, which assumes that the discrimination parameter $a_i$ is constant and equal to one (i.e., $\bar{a} = 1$) for all $i = 1, ..., k$. The Rasch model resulting from these assumptions is provided in Equation 13 and corresponds to Fischer and Molenaar’s formulation of the Rasch model (1995, p. 4).

$$P(x_{im} = 1|\theta_m) = \frac{\exp[D(\theta_m-b_i)]}{1+\exp[D(\theta_m-b_i)]}$$ (13)
Despite the mathematical equivalence of 1PL and Rasch models through appropriate rescaling, it is argued by some that Rasch modeling “represents a different philosophical perspective than that embodied in the 1PL model” (de Ayala, 2009, p. 19). According to de Ayala (2009, p. 19), Rasch modeling is used to construct the variable of interest and, consequently, the interested reader may refer to Fischer and Molenaar (1995) for a more comprehensive reference of Rasch modeling and of the various Rasch model extensions.

As seen by the model types presented above, the central idea of IRT is to relate the probability of a correct response on a particular item given a latent trait level as a function of one or more parameters that characterize the item (Fischer & Molenaar, 1995, p. 4). According to Fischer and Molenaar (1995, p. 4) as well as Baker (2001), the probability of a correct response to a particular item, \( i \), as a function of the latent trait level can be visualized on a graph and represents the item characteristic curve (ICC) or item response function (IRF). Each item on an instrument has a corresponding ICC. Figure 12 depicts the item characteristic curves for 25 items.

![Item Characteristic Curves](image)

**Figure 12.** Item characteristic curves for 25 dichotomous items.
(i.e., exam questions) from a dichotomously scored instrument, assuming equal discrimination for all items. IRT model parameters and how they relate to the ICC of an item will be discussed.

**Model parameters.** The most general formulation of an IRT model, presented in Equation 9, has three item parameters representing the item discrimination \( (a_i) \), the item difficulty \( (b_i) \), and the item pseudo-guessing \( (c_i) \). These item parameters and their relationship to the ICCs are explained below.

The item discrimination parameter \( (a_i) \) corresponds to the slope of the ICC. As such, higher values (in magnitude) for \( a_i \) correspond to steeper ICCs and suggest a more highly discriminating item whereas lower values (in magnitude) for \( a_i \) correspond to flatter ICCs and suggest a less discriminating item (Hambleton & Swaminathan, 1985, p. 38). While item discriminations could theoretically be negative values, this is unlikely in an educational testing context as this would suggest that examinees of lower latent ability levels have higher probabilities of getting a particular item correct. Assuming, then, that item discrimination parameters are nonnegative and that items could be appropriately reverse-coded as necessary before calculating discrimination parameters, items that have larger \( a_i \) values can more easily differentiate between examinees with differing ability levels whereas items with smaller \( a_i \) values cannot. Hambleton and Swaminathan note that the slope of the ICC at the point of inflection is related to the discrimination parameter as presented in Equation 14 (1985, p. 38).

\[
\text{Slope} = 0.425a_i(1 - c_i)
\]  

Equation 14

The item difficulty parameter \( (b_i) \) corresponds to the point on the ability scale where the slope of the ICC attains its maximum value (see Equation 14). This item difficulty parameter \( (b_i) \) then corresponds to the latent ability level where the probability of a correct response on item \( i \) is equal to \( \frac{1+c_i}{2} \) (Hambleton & Swaminathan, 1985, p. 38-39). In the 1PL and 2PL models, where
\( c_i = 0 \), this implies that when \( \theta = b_i \) the probability of correctly answering item \( i \) is 0.50. As such, higher values for \( b_i \) correspond to more difficult items whereas lower values for \( b_i \) correspond to easier items. Differences in \( b_i \) values for two items can be observed when visualizing the ICCs for each item. For instance, referring to Figure 12, the ICC for question 12 (represented as ‘Q12’) has a point of inflection that is at a lower value of the ability scale compared to question 17 (represented as ‘Q17’), which suggests that question 12 is an easier item than question 17; examinees at lower ability levels having a greater probability of responding correctly to question 12 than to question 17.

The item pseudo-guessing parameter \( (c_i) \) corresponds to the lower asymptote of the ICC and represents the probability that a low ability examinee (as \( \theta \to -\infty \)) will correctly answer the item (Hambleton & Swaminathan, p. 38-39). This pseudo-guessing parameter shifts the ICC vertically to account for any chance probability of a correct response on item \( i \) for an examinee with no ability and is assumed to be zero for all IRT model specifications except the 3PL IRT model (see Equation 10).

**Model parameter estimation methods.** Despite only discussing the model item parameters in the previous section, IRT models require estimation of both the item parameters for \( k \) items (corresponding to \( k \), \( 2k \), or \( 3k \) parameters for 1PL, 2PL, and 3PL models respectively) and the ability parameters for \( N \) examinees. Estimations of these item and ability parameters with both Maximum Likelihood and Bayesian estimation approaches are possible but seemingly complex given that to estimate the item parameters, the ability parameters must be specified and, similarly, to estimate the ability parameters, the item parameters must be specified.
Rizopoulous (2006) notes that estimation of IRT model parameters receives a great deal of attention in the literature. This attention paid to estimation procedures is in part due to the number of parameters that must be estimated and, also, to the apparent complexity of estimating parameters which are unobservable. In fact, Hambleton and Swaminathan (1985, p. 126-127) note that indeterminacy exists in the IRT model resulting in an identification problem, which may be removed by scaling the abilities (\(\theta_s\)) so that their mean is fixed. Additionally, unlike common statistical models, where the number of parameters is independent of the number of observations, the estimation of individual abilities results in a number of parameters that increases both with the number of items, \(k\), and with the number of examinees, \(N\). Hambleton and Swaminathan (1985, p. 129) point out that if the item and ability parameters are estimated simultaneously, estimators for item parameters (or ability parameters) are not consistent in the presence of infinitely many examinees (or items).

Common estimation procedures for unidimensional IRT item and ability parameters can be distinguished by the method by which ability parameters are estimated (whether they are jointly estimated with the item parameters, eliminated by conditioning, or integrated out by marginalization) and also by whether estimation is carried out using maximum likelihood (MLE) or some other method (Fischer & Molenaar, 1995, p. 40-41).

Baker and Kim (2004) describe various estimation techniques in detail including MLE estimation of item parameters with known ability parameters, MLE estimation of ability parameters with known item parameters, joint maximum likelihood estimation (JMLE), conditional maximum likelihood estimation (CMLE) that holds only in the 1PL/Rasch model case, and marginal maximum likelihood estimation (MMLE).
While Baker and Kim (2004) do provide some detail regarding Bayesian methods for estimating IRT model parameters such as expected a posteriori (EAP) and maximum a posteriori (MAP) strategies, most of these strategies are somewhat outdated. More contemporary Bayesian estimation techniques for IRT models relying on Markov chain Monte Carlo (MCMC) sampling methods, such as the Metropolis-Hastings (MH) within Gibbs sampling algorithm proposed by Patz and Junker in 1999 are described in detail in Fox (2010).

Given that the goal of this study is not to explore nor to make comparisons about the various IRT estimation methods available, only a few notes are currently made about the availability of different estimation methods. In addition, the estimation methods employed in the current study, namely MMLE with maximization of the integrated log-likelihood with respect to $\theta$ achieved using the BFGS algorithm (a quasi-Newton optimization method named after Broyden, Fletcher, Goldfarb, and Shanno), will be discussed in Chapter 3.

For detailed descriptions of item and ability parameter estimation techniques from frequentist and Bayesian perspectives as well as the relative merits of each method, the interested reader is referred to de Ayala (2009), Baker and Kim (2004), and Fox (2010). Additionally, Nocedal and Wright (1999) is a comprehensive reference for details regarding various Newton and quasi-Newton methods for numerical optimization.

**Model assumptions.** Unidimensional IRT models, as described above, have a number of assumptions. One assumption underlying the model is that of unidimensionality. As described by de Ayala (2009, p. 20), unidimensionality “states that the observations on the manifest variables (e.g., the items) are solely a function of a single continuous latent person variable.” This unidimensionality assumption could be formulated as in Equation 15, which demonstrates that the probability of a particular response for item $i$ depends only on the ability of an individual and
on the item parameters specified by the model \((a_i, b_i, \text{ and } c_i)\) in the most general 3PL IRT model but not on any other variable, \(\varphi\).

\[
P(x_{im} = 1|\theta_m, a_i, b_i, c_i, \varphi) = P(x_{im} = 1|\theta_m, a_i, b_i, c_i)
\]

(15)

The unidimensionality assumption is considered an ideal situation in that, as noted by de Ayala (2009, p. 20), violations of unidimensionality may or may not be problematic since the unidimensional model might sufficiently represent data that is in truth a manifestation of two or more latent traits. However, as noted by Hambleton and Swaminathan (1985, p. 156), unidimensionality is desirable for test construction as it enhances test score interpretability.

A second assumption of unidimensional IRT models is conditional or local independence. Local independence is related to the concept of unidimensionality and states that the response to a particular item is solely determined by an individual’s location on the latent trait continuum (de Ayala, 2009, p. 20). Consequently, item responses are independent given an individual’s latent trait value and can be formalized as in Equation 16.

\[
x_{im} \perp x_{jm} | \theta_m \ \forall \ i \neq j
\]

(16)

The property of local independence is sufficient for meeting the assumption of unidimensionality (Hambleton & Swaminathan, 1985). However, Goldstein found that the unidimensionality assumption and the local independence assumption are not always the same (as cited in de Ayala, 2009, p. 21). Local independence does provide for the easy calculation of the probability of a particular item response string given a known ability level, as this becomes the product of the probabilities for responses on individual items.

A third assumption of unidimensional IRT models is the functional form assumption and is an assumption regarding the nature of the ICC. This assumption states that the data follow the
form of the ICC specified by the stated IRT model (de Ayala, 2009, p. 21). This includes
assuming the monotonicity of the IRF so that for every $\theta_u, \theta_w$ where $\theta_u > \theta_w$:

$$P(x_{iu} = 1|\theta_u, a_i, b_i, c_i) > P(x_{iw} = 1|\theta_w, a_i, b_i, c_i)$$  (17)

An additional assumption of unidimensional IRT models is parameter invariance stating that item parameter estimates do not depend on the sample of examinees and should be equivalent up to some linear constant (Hambleton & Swaminathan, 1985, p. 155-169).

With the goal of developing a localized IRT modeling technique that allows for the investigation of spatially varying item functionality of survey instruments and educational assessments, especially in the context of ILSAs, a brief overview of IRT models has been provided and, now, a non-exhaustive background on item functionality and DIF detection techniques will also be provided.

**Differential Item Functioning**

Differences in individual item functionality resulting from the presence of nuisance determinants, which are abilities that influence the response of an examinee on a particular item but are not the target ability the instrument was intended to measure, can combine to create a “coherent and major biasing influence at the test level” even when the item-level DIF is small in magnitude (Shealy & Stout, 1991). If item-level DIF can combine to create test-level bias, the interpretation of ILSA score comparisons between groups, made on cross-national or subnational levels, are limited. Svetina and Rutkowski (2014) note that for scores to be truly comparable across groups or subpopulations, measurement invariance must hold. Differences in item functionality, then, are of concern not only in educational contexts but, also, in any high stakes context where inappropriate decisions regarding group or subpopulation comparisons can have extremely adverse effects such as those decisions made in medical screening tests (see Longford,
The effect of DIF can be measurement bias or even discrimination (Tutz & Berger, 2016). It is quite possible that items on survey instruments may function differently across space due to regional disparities that might be attributable to a number of factors or to an intricate interaction of several factors. In order to provide adequate background for the investigation of regional DIF detection, working definitions for DIF are provided as well as a brief discussion of select traditional and new DIF detection techniques.

Definitions. Differential item functioning (DIF) occurs when items function differently for individuals of the same latent ability or skill level based upon a class or group membership. According to Zumbo (1999), DIF occurs when the probability of responding correctly to an item differs due to group membership despite individuals having the same latent ability level. DIF can also be thought of as a systematic error in how an item measures a latent construct for members of a particular group (Camilli & Shepard, 1994). Ackerman (1992) notes that DIF occurs only when an item unintentionally measures more than one latent trait (i.e., the item measures both a primary and a secondary dimension) and when groups have different ability distributions on the secondary dimension.

Throughout the literature, there is a distinction between item bias and item impact. According to Clauser & Mazor (1998), item bias occurs when examinees across groups respond to an item differently because of differences on an invalid construct that the item was not designed to measure (i.e., a secondary or nuisance dimension) whereas item impact occurs when examinees across groups respond to an item differently because of differences on a valid construct that the item was designed to measure (i.e., the primary or target dimension).

While DIF broadly implies that an item performs “differently for one group of examinees relative to the way it performs for another group of examinees,” two specific types of DIF
(uniform and non-uniform) are recognized (Zwitser & Glaser, 2017). According to Zwitser and Glaser (2017), uniform DIF suggests that an item is uniformly or systematically more difficult for one group of examinees than another when the groups of examinees are matched on ability. An example of uniform DIF that favors the reference group at all ability levels is provided in Figure 13. It is of note that uniform DIF can be seen as a shift in item difficulty parameters between two compared groups (reference and focal). According to Zwitser and Glaser (2017, p. 211), non-uniform DIF “means that the correlation between a particular item response and the latent variable varies across subpopulations.” Non-uniform DIF suggests that an item is more difficult for one group of examinees than another when the groups of examinees are matched on ability at one end of the continuum and less difficult for this group of examinees at the other end of the continuum, after both groups have been matched on ability. Essentially, the shift in item difficulty is not consistent at all ability levels and the favored group differs based upon where on

![Graph](https://via.placeholder.com/150)

*Figure 13. Uniform differential item functioning favoring the reference group.*
the ability continuum groups of examinees are matched. De Ayala (2009, p. 343) summarizes non-uniform DIF as an interaction between item performance, group membership, and latent ability level. An example of non-uniform DIF that favors the reference group at most ability levels but favors the focal group at lower ability levels is provided in Figure 14. Non-uniform DIF can be seen as a result of differing discrimination parameters between two compared groups (reference and focal).

![Figure 14. Non-uniform differential item functioning favoring the focal group at lower ability levels but favoring the reference group at higher ability levels.](image)

 Observable group differences in ICCs suggest the presence of DIF. However, it is worthwhile to note that group differences in average abilities should not affect the estimated ICCs, as these curves will match the two groups on ability.

**Traditional differential item functioning detection methods.** While a more thorough overview of existing DIF detection methods (both traditional and contemporary) can be found in Magis, Beland, Tuerlinckx, and Boeck (2010), the current work will provide descriptions of
three traditional methods that will be utilized in the study: Mantel-Haenszel chi-square, logistic regression, and Raju’s area methods.

Traditional DIF detection techniques such as these three typically consider only two groups (reference and focal). According to de Ayala (2009, p. 343), the focal group is the group investigated for disadvantage (or advantage) on a particular item and is, often, the “minority” group whereas the reference group serves as the comparison group and is, often, the “majority” group. Traditional DIF detection procedures include contingency table methods, logistic regression methods, IRT based methods, and structural equation modeling (SEM) methods. The three methods presented here represent each of the DIF detection methods except SEM.

**Mantel-Haenszel chi-square.** The Mantel-Haenszel (M-H) chi-square method presented by Mantel & Haenszel in 1959 and detailed by Holland and Thayer (1988) is a contingency table method that can be used for DIF detection where dichotomous item response and group membership are assessed for independence while conditioning on the observed raw test score (as cited in de Ayala, 2009, p. 327). Essentially, the M-H chi-square method equates to the analysis of a three-way contingency table. In addition to testing the null hypothesis of conditional independence using the M-H chi-square method, an indication of the odds ratio of success on a particular item for the reference group members compared to the focal group members is denoted as $\hat{\alpha}_{MH}$ and can be calculated as in Equation 18 when the 2 x 2 table of manifest groups by item response can be formed for the $j$th observed raw test score as in Table 2 where $A_t, B_t, C_t, D_t, n_{Rt}, n_{Ft}, m_{0t}, m_{1t},$ and $T_t$ represent frequencies for the corresponding cells and $L$ represents the instrument length:

$$\hat{\alpha}_{MH} = \frac{\sum_{t=1}^{(L-1)} A_t D_t}{\sum_{t=1}^{(L-1)} B_t C_t} \frac{T_t}{T_t}$$

\[ (18) \]
Table 2. *Mantel-Haenszel contingency table conditioned on the jth raw test score.*

<table>
<thead>
<tr>
<th>Item Response</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference</strong></td>
<td>$B_t$</td>
<td>$A_t$</td>
<td>$n_{Rt}$</td>
</tr>
<tr>
<td><strong>Focal</strong></td>
<td>$D_t$</td>
<td>$C_t$</td>
<td>$n_{Ft}$</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>$m_{0t}$</td>
<td>$m_{1t}$</td>
<td>$T_t$</td>
</tr>
</tbody>
</table>

**Logistic regression.** The logistic regression method can be used for DIF detection where a logistic regression model is performed to predict the outcome of a dichotomous item response conditioned on the observed raw test score in two separate analyses (de Ayala, 2009, p. 332). The first analysis uses members of the focal group for calibration while the second analysis uses members of the reference group for calibration. The estimated $\beta_0$ and $\beta_1$ coefficients are compared for reference and focal groups, allowing for the identification of both uniform and non-uniform DIF (de Ayala, 2009, p. 332). While this logistic regression strategy can be viewed as two separate analyses, the implementation of this method for DIF detection actually consists of fitting one logistic model with the observed raw test score, the group membership, and an interaction between both as covariates (Magis, Beland & Raiche, 2016). The statistical significance of the parameters related to group membership and the group-score interaction is then evaluated by means of either the likelihood-ratio test or the Wald test to identify uniform and non-uniform DIF (Magis et al., 2016).

**Raju’s area.** IRT based DIF detection methods involve either comparing item parameter estimates, under the assumption of parameter invariance, or comparing ICCs using area methods.
Raju’s method (Raju, 1988; Raju, 1990) is an IRT based area method of DIF detection where probabilities of correct item responses are compared across groups conditioned on latent ability estimates as specified by an IRT model. The implementation of this method for DIF detection actually involves finding the unsigned, or signed, area between two ICCs that correspond to the two comparison groups and are estimated using an appropriate IRT model (Magis, Beland & Raiche, 2016). This method allows for the identification of both uniform and non-uniform DIF (Magis et al., 2016).

**New differential item functioning detection methods.** Traditional DIF detection methods are based on test statistics and focus on a priori subgroups (Tutz & Berger, 2016). The limitation of traditional techniques to only a few subgroups, often two (reference and focal) necessarily assumes that these two group classifications are meaningful and that the individuals within each manifest group are homogenous (de Ayala, 2009, p. 407; Tutz & Berger, 2016). Consequently, newer DIF detection methods try to provide techniques that are applicable to multiple group settings. Initial attempts to detect DIF with multiple groups typically focused on the application of several time consuming and costly pairwise comparisons (Hambleton & Kanjee, 1995; Svetina & Rutkowski, 2014). Moreover, many of these approaches amount to artificially creating two groups, one group against the aggregate of the other groups or against the composite of all groups. Some of these methods for DIF detection in a multiple group context include the OECD ANOVA-like procedure for examining item-by-country interactions, extensions of CFA and IRT to a multiple group setting, MIMIC and MIMIC-interaction models, LR Lasso DIF, the Alignment Method, and recursive partitioning approaches such as Rasch trees and item-focused trees.
Many of the aforementioned methods still require a priori, pre-specified, grouping variables. Methods that do not require a priori subgroups, such as Rasch tree methods, still produce results that are within the scope of pre-specified covariates (Apinyapibal et al., 2015). Additionally, Rasch tree models suggested by Stobl et al. (2010) do not actually identify items responsible for DIF but rather regions of the covariate space that are linked to DIF (Tutz & Berger, 2016). Similarly, MIMIC models do not identify item-level DIF (Finch, 2005).

While contemporary DIF detection techniques are continually developing and expanding to handle issues such as the comparison of multiple groups, the accommodation of continuous variable factors, and avoidance of a priori specification of subgroups, none of the methods currently provided above recognize that demographic and political borders that currently define spatial comparison subgroups are arbitrary and artificial. As such, the utilization of demographic and political boundaries for subgroup definitions in multiple group comparisons assumes homogenous manifest groups, suffers from the MAUP issues that plague global spatial models, and ignores the conceivable spatial structure of the observations. Additionally, despite the beneficial visualization provided by the newly developed recursive partitioning tree based approaches to DIF detection, the visualization offered by mappings of estimated local IRT model parameters may have great potential for policy-oriented research, especially in the context of ILSAs that serve as change agents driving reform across and within educational systems (Braun, 2013; Ritzen, 2013).

Consequently, there is a need to propose a localized IRT model for the investigation of potential regional DIF based upon geographic location without the need for pre-specified groupings and before any confirmatory DIF testing. The proposed methodology for item-level
regional DIF detection will be outlined in the following chapter as will the various simulated and empirical case studies, which will illustrate the methodology.
Chapter 3

Method

The primary objective of the current study is to propose a localized approach to IRT modeling that can detect item-level regional DIF based upon geographic location without the need for pre-specified groupings and before any confirmatory DIF testing. Beyond proposing a localized IRT model that can account for spatial structure inherent in large scale assessments, a secondary objective of the current study is to illustrate the proposed method utilizing simulated case studies and one empirical application. While all of the case studies provided are intended to demonstrate certain sensitivities of the proposed method to procedural choices and circumstances such as bandwidth selection, model choice (i.e., 1PL or 2PL), and sample size variation, each individual case study serves additionally to explore potential drawbacks and/or benefits of the proposed method given changes in observed DIF type, DIF magnitude, and spatial structure.

The localized approach to IRT modeling as well as the step-by-step method for item-level regional DIF detection will be outlined. Following this outline, the data and each of the various case studies will be described.

Local Item Response Theory

It is quite possible that items on large scale assessments, especially educational assessments where national and subnational comparisons are intended, function differently due to regional or spatial location. Detection of and further investigation into such regional DIF in educational settings may allow for benchmarking and increased educational opportunity, serving as an agent of change in educational policy. Despite ongoing attempts to consider DIF across spatial subgroups defined by political boundaries, no method has yet to truly consider and account for the underlying spatial structure of observations. Local IRT modeling, assuming
spatial nonstationarity, utilizes a moving window approach to IRT to provide disaggregated decompositions of the global model.

**Model.** Utilizing a moving window technique, local IRT models account for spatial structure and address nonstationarity directly by allowing item parameters to vary across space. Local IRT models are calibrated separately at several regression points that sweep across the spatial region of interest. Item parameter estimates, consequently, are location dependent with individual examinee data contributing to each local IRT model according to a box-car kernel binary weighting scheme and a discrete zoning system with a fixed or an adaptive bandwidth.

Suppose that $x_{im}$ is the dichotomous response on the $i$th item of an instrument for the $m$th examinee with a corresponding latent ability or skill level of $\theta_m$ and that each of the $m$ examinees can be associated with a spatial position available in a suitable coordinate system so that the $j$th point in space has coordinates denoted by $(u_j, v_j)$. Recall that in IRT modeling a scaling factor of $D = 1.702$ may be used to equate, approximately, the normal and logistic metrics. Further, recall that values of $a_i$, $b_i$, and $c_i$ represent the discrimination, difficulty, and pseudo-guessing parameters for the $i$th item of an instrument respectively. By incorporating the scaling factor, $D = 1.702$, and assuming a logit link function, the global unidimensional IRT model formulated as in Equation 10, which corresponds to Hambleton and Swaminathan’s formulation of the 3PL IRT model (1985, p. 37-38), can be localized with the following form specific to location $u$, where $u$ is a vector of the coordinates $(u_j, v_j)$:

$$
P(x_{im} = 1 | \theta_m) = c_i(u) + \left(1 - c_i(u)\right) \frac{\exp\{D a_i(u)(\theta_m - b_i(u))\}}{1 + \exp\{D a_i(u)(\theta_m - b_i(u))\}}
$$

(19)

It should be noted that a local IRT model could be modified as a mixed model wherein some of the parameter estimates are fixed (i.e., they do not vary over space).
Despite the specification of the more general 3PL local IRT model provided in Equation 19, the current study will only incorporate 2PL and 1PL local IRT models. The rationale for utilizing only 2PL and 1PL models in the local IRT modeling framework is based upon the observed numeric instability of the 3PL global IRT model and the large calibration sample sizes required, which might not be practically available at a local level (de Ayala, 2009, p. 131).

The reduction of Equation 19 to the 2PL local IRT model provides added numeric stability to the local calibrations and lowers the calibration sample sizes required at each regression point while still providing for a spatially-varying discrimination parameter, \( a_i(u) \), which might be utilized for detection of non-uniform regional DIF. In the 2PL local IRT model, there is no pseudo-guessing parameter i.e., \( c_i(u) = 0 \). The resulting local IRT model is provided in Equation 20.

\[
P(x_{im} = 1|\theta_m) = \frac{\exp\{Da_i(u)(\theta_m-b_i(u))\}}{1+\exp\{Da_i(u)(\theta_m-b_i(u))\}}
\]  

Despite the added numeric stability to the local calibrations that comes from a 2PL local IRT model as well as the ability to detect non-uniform regional DIF through the analysis of spatial nonstationarity in the discrimination parameter, the calibration sample size required for a 2PL IRT model is at least 500 persons in even ideal conditions (de Ayala, 2009, p. 105). Samples of this size may still not be practically available at a local level. As such, Equation 20 can be further reduced to a 1PL local IRT model by assuming there is no pseudo-guessing parameter and by holding the spatially-varying discrimination parameter, \( a_i(u) \), constant for all instrument items \( i = 1, \ldots, k \). This suggests that the discrimination parameter would be constant across items, specific to a location, but would vary across locations without constraint. However, the discrimination parameter could be constant across items and fixed across space. This restriction as well as the assumption of a unit discrimination parameter as in the Rasch model will be
assumed for the 1PL local IRT model in the context of this study. Consequently, the spatially-varying discrimination parameter \( a_i(u) \) of Equation 20 is held constant across items and across space and is set equal to one (i.e., \( a_i(u) = 1 \)) for all \( i = 1, ..., k \) and all locations, \( u \). The IRT model resulting from these assumptions is provided in Equation 21.

\[
P(x_{im} = 1|\theta_m) = \frac{\exp[D(\theta_m - b_i(u))]}{1+\exp[D(\theta_m - b_i(u))]} \tag{21}
\]

The moving window technique to IRT modeling described in this study involves the repeated calibration of the selected IRT model, each using only a subset of observations that lie within a circular spatial region. The specific method of local IRT model calibration and parameter estimation will be described below.

**Model calibration.** The local IRT models provided above represent several IRT models that are location dependent. The moving window approach to IRT incorporates spatial structure and allows for spatial changes in the estimated item parameters to be monitored and subsequently mapped for visualization. The defined spatial regions can be modified through the specification of regression points (i.e., the center of the moving windows). Moreover, assuming circular moving windows, the calibration sample sizes and the area of the spatial regions can be modified through the specification of a fixed or an adaptive bandwidth corresponding to the radius of the circular moving windows.

It is important to note that the moving window approach to local IRT modeling is equivalent to a box-car kernel binary weighting scheme on examinee observations and a discrete zoning system with a fixed or an adaptive bandwidth. While a different weighting scheme and a fuzzy zoning system may seem of interest for modeling purposes, the results are assumedly relatively insensitive to the weighting function as noted in the context of GW modeling (Fotheringham et al., 2002, p. 44). In addition, according to Gollini et al. (2015), this box-car
kernel, moving window approach, is more computationally efficient when dealing with large datasets. Given that the purpose of this proposed local IRT method is to detect regional DIF across several spatial locations in the context of ILSAs, large datasets on a global level are presumed. Consequently, local IRT models in this study will be calibrated at regression points, designated to be the centroids of each areal unit, using a box-car kernel weighting scheme and a fixed bandwidth.

Given that results are most influenced by and most sensitive to the choice of the bandwidth (Charlton & Fotheringham, 2009), several user-specified, fixed bandwidths will be employed for each case study. Fixed bandwidths will be employed since the purpose of the proposed method is to detect regional DIF across several regions and, potentially, to reduce the number of pairwise comparisons necessary in confirmatory DIF testing. Since the groups that will be utilized in confirmatory DIF testing are data-driven (i.e., they are not specified a priori) and since all simulated local sample sizes will be set in this study to levels that would not be considered sparse, adaptive bandwidths to equalize calibration sample sizes will not be considered at this time. User-specified bandwidths will be employed because, as suggested by Fotheringham et al. (2002, p. 158), this allows for local results to be directly investigated for sensitivity to spatial scale. In GW modeling, an ‘optimal’ bandwidth can be found by optimizing the cross-validation (CV) score or some other criterion measure such as AIC, AICc, or BIC (Fotheringham et al., 2002, p. 60-62). However, every bandwidth selection results in a variance-bias tradeoff with larger bandwidths providing only broad details of spatial nonstationarity and smaller bandwidths reducing the calibration sample size to points at which results become unstable (Fotheringham et al., 2002, p. 143-144). Additionally, local IRT modeling is a new approach and, as such, the sensitivity of local IRT model results to bandwidth selection has not
been investigated. Consequently, the local IRT models in this study will be calibrated at each areal unit centroid using a box-car kernel weighting scheme and several user-specified, fixed bandwidths.

The calibration samples, then, are local subsets of the data that are within a fixed distance from the specified regression points. An observation is included in a given local model calibration if the distance between the observation and the regression point is less than the bandwidth, \( h \). Note that \( h \) will be used to represent the bandwidth from this point in the study onward in order to avoid confusion with the spatially-varying local IRT difficulty parameter for the \( i \)th item, \( b_i(\mathbf{u}) \).

Each local IRT model will be fit using MMLE. As noted by Rizopoulos (2006), MMLE estimates model parameters by maximizing the observed data log-likelihood that is obtained by integrating out the latent variables. Essentially, item parameters are estimated by maximizing the marginal likelihood function (de Ayala, 2009, p. 70). The contribution of the \( m \)th examinee to the integrated log-likelihood to be maximized is provided in Equation 22 (Rizopoulos, 2006)

\[
\ell_m(a_i, b_i) = \log \int p(x_m | \theta_m, a_i, b_i) p(\theta_m) d\theta_m
\]  

(22)

where \( p(\cdot) \) denotes a probability density function, \( x_m \) denotes the vector of responses for the \( m \)th examinee, and \( \theta_m \) is assumed to follow a standard normal distribution. The integral provided in Equation 22 is approximated with the Guass-Hermite quadrature rule and 21 quadrature points. The integrated log-likelihood is then maximized with respect to \( \theta \) using the Broyden-Fletcher-Goldfarg-Shanno (BFGS) quasi-Newton optimization method (Rizopoulos, 2006). This estimation procedure is similar to that described by de Ayala (2009, p. 68-79, 356-359) and, also, by Baker and Kim (2004, p. 157-176) but utilizes the BFGS quasi-Newton optimization method as opposed to the EM algorithm, which might be slow to converge in practice. Refer to Nocedal
and Wright (1999) for further details regarding Newton and quasi-Newton methods for numerical optimization. As Baker and Kim observe (2004, p. 175), an additional estimation procedure must be paired with MMLE/BFGS to estimate individual examinee abilities, such as MLE, expected a posteriori (EAP), or maximum a posteriori (MAP). Despite this inconvenience with MMLE and although JMLE was at one time the standard estimation method for IRT, Baker and Kim (2004, p. 94-107, 175) note several problems with JMLE and comment on the advantages of MMLE. For instance, JMLE suffers from the Neyman-Scott problem wherein JMLE item parameter estimates are not consistent in the presence of the ability parameters (Baker & Kim, 2004, p. 175). MMLE does not suffer from this problem and provides consistent item parameter estimates. With the goal of detecting regional DIF by observing nonstationarity in item parameter estimates across several model calibrations, each local IRT model will be fit using MMLE.

In summary, local IRT models will be fit using MMLE and will be calibrated at each areal unit centroid using a box-car kernel weighting scheme with several fixed bandwidths. All local IRT model calibrations will be performed in R (R Development Core Team, 2017; Rizopoulos, 2006, 2015).

**Detection of spatially varying differential item functioning.** As previously noted, uniform DIF can be seen as a shift in item difficulty parameters between two compared groups (reference and focal) whereas non-uniform DIF can be seen as a difference in item discrimination parameters between two compared groups (reference and focal) when the groups of examinees are matched on ability. Consequently, observable group differences in item parameters, when groups are matched on ability, can suggest the presence of DIF, both uniform and non-uniform. These observable differences in item parameters would be reflected in the
group item characteristic curves, ICCs, which would provide visual evidence supporting the presence of potential DIF.

Using the idea that violations of IRT assumptions (such as parameter invariance) across identifiable spatial or regional groups matched within ability distribution provides us with a working definition of regional DIF in space, mappings of spatially-varying local IRT item parameters (difficulty and/or discrimination) will allow for visual investigation of potential DIF based upon geographical location without need for pre-specified groupings and prior to any confirmatory DIF testing. This visual investigation of spatial trends in item parameter estimates across a spatial region of interest could allow for meaningful spatial subgroups to be created, thereby minimizing the number of pairwise comparisons needed in confirmatory DIF testing. The creation of spatial subgroups can be accomplished by treating the local IRT item parameter estimates as fixed when calculating location-specific measures of spatial autocorrelation. The extent to which the spatially-varying local IRT item parameter estimates are more similar or more dissimilar when close to a specific location in space can be assessed using a LISA such as Local Moran’s I (Anselin, 1995). Local Moran’s I values will provide a way to identify statistically significant local effects (e.g., clusters and hotspots) in the item parameter estimates and, correspondingly, to assign areal units to larger spatial subgroups if desired and applicable, which might be used in subsequent DIF analyses.

**Procedural Steps for the Proposed Method**

To summarize the proposed method of local IRT for the detection of spatially-varying DIF, an outline of the step-by-step procedure is provided below.
1. Select the location for regression points, \( \mathbf{u} \), at which local IRT models will be calibrated. Recall that \( \mathbf{u} \) is a vector of coordinates \((u_j, v_j)\) that need not be the points where data were collected.

2. Select an appropriate distance metric so that the distance, denoted \( d_{jk} \), between any two spatial locations, \( j \) and \( k \), may be defined.

3. Select a fixed or adaptive bandwidth, \( h \), which will create a neighborhood for each regression point location and which, in a moving window approach, will define the local subsets of data serving as calibration samples.

4. Utilize a moving window approach to fit local IRT models (1PL or 2PL) at each regression point location, \((u_j, v_j)\). Parameter estimation for these local IRT models will be achieved using MMLE. Note that, if equal ability distributions cannot be assumed across the regional subgroups, an anchor set of items can be selected and utilized in the local IRT calibrations.

5. Map the estimated item parameters as well as the associated standard errors for visual investigation of potential regional DIF and spatial nonstationarity of item parameters.

6. Identify spatial subgroups for further investigation and subsequent DIF analyses using calculated Local Moran’s I values and corresponding significance testing procedures. Mappings of the Local Moran’s I values can also be made to reveal stronger or lesser local spatial autocorrelation in item difficulty and discrimination parameter estimates across the spatial region of interest (Mennis, 2006).

7. Using the potentially reduced number of spatial subgroups, perform a confirmatory DIF detection procedure if desired. Ideally, the method will reduce the number of spatial subgroups, minimizing the number of pairwise comparisons necessary in
traditional DIF detection techniques such as the Mantel-Haenszel (M-H) Chi-square, logistic regression, and Raju’s area methods. Even if new DIF detection techniques for multiple groups will be utilized, the spatial subgroups identified by the proposed local IRT modeling approach may now serve as more meaningful a priori group specifications in subsequent confirmatory DIF analyses, as these groupings were obtained in a data-driven way incorporating spatial structure. It is worth noting that if no identifiable spatial subgroups emerge from the local IRT modeling approach, DIF attributable to spatial location may not be present.

**Simulated Data**

The proposed methodology, outlined above, will be demonstrated with case study applications to simulated data as well as an application to an empirical dataset. The simulation data process and the simulated case study factors are described below.

**Data generation.** WinGen3 (Han, 2007) will be used to simulate examinee response data to 25 dichotomously scored items at separate locations over one spatial region. The spatial region utilized for the simulated case studies is a 75 areal unit irregular lattice in the geographic coordinate system (i.e., longitude, latitude coordinates).

**Non-manipulated factors.** In order to describe and illustrate the methodology practically by providing detailed examples of several simulated case studies, with comparisons made to traditional DIF techniques, select factors will be held constant. The moving window approach to local IRT modeling, utilizing a box-car kernel with a discrete zoning system, will be held constant throughout the study. For computational efficiency in the context of large global datasets, no other weighting schemes will be attempted in the current study.
In addition, adaptive bandwidths will not be considered in the current study. Only fixed bandwidths will be employed. However, three fixed bandwidths will be utilized in each simulated data case study to demonstrate the sensitivity of local IRT results to bandwidth selection, a demonstration that Fotheringham et al. (2002, p. 158) consider invaluable. Consistently utilizing a fixed bandwidth in the demonstrations, rather than an adaptive bandwidth, will also allow for the investigation of model issues related to observable differences in calibration sample sizes, especially as the local sample sizes are manipulated.

The last factor that will not be manipulated or investigated in the simulated case studies is instrument length. Also known as test length, instrument length will be held constant throughout the demonstrations of the local IRT method for regional DIF detection. Several DIF simulation studies have utilized instrument lengths between 20 and 60 items (Chang, Mazzeo & Roussos, 1996; Fidalgo, Mellenbergh & Muniz, 2000; Finch & French, 2008; French & Maller, 2007; Hidalgo-Montesinos & Lopez-Pina, 2002; Kim & Cohen, 1992). Moreover, de Ayala (2009, p. 104-105) suggests that instruments of 20 or more items appear to provide reasonably accurate MMLE item parameter estimates for 2PL IRT models when calibration sample sizes are sufficiently large and when other modeling conditions are favorable. Consequently, instrument length for the simulation data will be fixed at 25 items in the current study.

**Manipulated factors.** The simulated case studies serve as detailed examples to illustrate the use of the proposed method, while investigating potential local IRT model sensitivities to spatial structures, types of DIF, magnitudes of DIF, bandwidth selections, local sample sizes, and IRT model choices. In addition, each individual case study is provided to specifically demonstrate hypothesized benefits while, also, discovering potential drawbacks of the proposed method given modifications of several factors.
A factor that will be slightly manipulated in the simulated studies but that will be further investigated in the context of the current study, owing to the application of the method to an empirical dataset, is the underlying spatial structure of the data. An irregular lattice with 75 areal units in geographic coordinate space will be utilized for the three simulated case studies. The irregular lattice used in these simulated case studies will be the state of Arkansas with the 75 areal units representing the 75 counties in the state. Though this is an irregular lattice, Arkansas is fairly regular with several counties being approximately rectangular in shape and of nearly equivalent area. Consequently, a less ideal spatial structure will be explored in the empirical application of the method to an irregular lattice with 28 areal units in geographic coordinate space. In the simulated case studies, the spatial structure of the data is manipulated by altering the placement of subgroups exhibiting DIF in the study region. The first case study will have only one latent spatial subgroup, indicating that there is no regional DIF present. Other simulated case studies will have two or a randomly assigned number of latent spatial subgroups. The number of latent spatial subgroups in the study region is manipulated to demonstrate the ability of the proposed method to correctly detect the number of latent spatial subgroups and to correctly identify which areal units should be included in these spatial subgroups. This spatial subgroup identification is important if subsequent, confirmatory DIF analyses of spatial subgroupings is a goal of the researcher.

Not only does the manipulation of the spatial structure demonstrate the proposed method’s ability to detect and identify spatial subgroups for subsequent, confirmatory DIF analyses but it also allows for the demonstration of certain hypothesized benefits of the method. For example, a local approach to IRT for regional DIF detection is proposed because it is hypothesized to reveal DIF that is unobservable on a global level and that is potentially
undetectable with a non-proximal approach (i.e., where traditional DIF techniques are applied to ‘spatial’ groups defined by cardinal direction only). Consider, for example, a regular lattice with 36 equally spaced areal units as seen in Figure 15. Traditionally, DIF testing using a non-proximal approach with spatial groups defined using only cardinal direction as in Figure 15 might be applied, comparing spatial subgroups such as that of North to South or East to West.

![Regular Lattice](image1)

*Figure 15. Regular lattice with subgroups defined by cardinal direction.*

However, consider the possibility that two latent spatial subgroups exist as seen in Figure 16. Consider, also, that there exists an item that functions differently on each of these two spatial subgroups. It is of note then that traditional DIF techniques comparing regions North to South or East to West would result in a spatial Simpson’s paradox where regional DIF would be unobservable. The potential to prevent a spatial Simpson’s paradox is just one of the hypothesized benefits of the proposed local approach to IRT modeling for regional DIF detection when comparisons are made to traditional DIF techniques using a non-proximal approach.

Another hypothesized benefit of the proposed local approach to IRT modeling for regional DIF
detection is to reveal spatial trends in DIF that would be undetectable on a global level. For example, consider the possibility that three latent spatial subgroups exist as seen in Figure 16.

Figure 16. Regular lattice with two or three latent spatial subgroups.

Consider, also, that there exists an item that functions differently on each of these three spatial subgroups such that an areal unit in one of the latent subgroups, when compared to an areal unit in an adjacent latent subgroup, would have a significant area between their two respective IRFs. Notice that this implies the area between respective IRFs of areal units located in the two non-adjacent latent subgroups would be even larger in magnitude. While a traditional, non-proximal approach to DIF detection would detect differences in comparisons made East to West, the magnitude of the regional disparities in item functionality and the spatial trend of those regional disparities would be unobservable without utilizing a local approach. The second simulated case study will involve manipulation of the spatial structure to demonstrate one of these hypothesized benefits.
DIF type and DIF magnitude will also be manipulated in these simulated case studies. One case study will involve no regional DIF in order to discern the impact of spatial structure, bandwidth selection, local sample size, and model type on false positive rates. Two of the simulated case studies, however, will involve uniform and non-uniform DIF. The magnitude of DIF will be modified by changing the area between spatial subgroup IRFs according to Raju’s area formula. Raju (1988, 1990) provided an exact unsigned area (EUA) formula to describe the area between reference and focal group IRFs for the 3PL IRT model. Raju’s EUA formula is presented in Equation 23.

\[
\begin{cases}
(1 - c) \left| \hat{b}_{jF} - \hat{b}_{jR} \right| \\
(1 - c) \left| \frac{2(\hat{a}_{jF} - \hat{a}_{jR})}{D\hat{a}_{jR}\hat{a}_{jF}} \ln \left( 1 + \exp \left[ \frac{D\hat{a}_{jR}\hat{a}_{jF}(\hat{b}_{jF} - \hat{b}_{jR})}{\hat{a}_{jF} - \hat{a}_{jR}} \right] \right) - (\hat{b}_{jF} - \hat{b}_{jR}) \right| 
\end{cases}
\]

\[\text{if } \hat{a}_{jR} = \hat{a}_{jF} \]
\[\text{if } \hat{a}_{jR} \neq \hat{a}_{jF} \]

Notice that, in the case of the 2PL IRT model, \( c = 0 \) and Equation 23 simplifies to the EUA formula that will be used in the context of this study and is provided in Equation 24 below.

\[
\begin{cases}
\left| \hat{b}_{jF} - \hat{b}_{jR} \right| \\
\frac{2(\hat{a}_{jF} - \hat{a}_{jR})}{D\hat{a}_{jR}\hat{a}_{jF}} \ln \left( 1 + \exp \left[ \frac{D\hat{a}_{jR}\hat{a}_{jF}(\hat{b}_{jF} - \hat{b}_{jR})}{\hat{a}_{jF} - \hat{a}_{jR}} \right] \right) - (\hat{b}_{jF} - \hat{b}_{jR}) 
\end{cases}
\]

\[\text{if } \hat{a}_{jR} = \hat{a}_{jF} \]
\[\text{if } \hat{a}_{jR} \neq \hat{a}_{jF} \]

Several DIF studies alter the magnitude of DIF for a particular item by manipulating the discrimination and/or the difficulty parameters for particular items, according to Equation 24, so that the area between IRFs is fixed to be approximately 0.40, 0.60, and 0.80 (Finch & French, 2008; French & Maller, 2007; Swaminathan & Rogers, 1990). Manipulating the type and the magnitude of DIF present between the spatial subgroups will allow for an assessment of whether the proposed method is powerful in detecting small levels of DIF or, owing to the disaggregation of the data, whether the proposed method is overly sensitive and provides false positive results.

The local sample size will also be a manipulated factor for each case study to determine its effect on the proposed method. Sample sizes and, consequently, sample size ratios for the
areal units are selected to represent relatively small ($n = 250$), moderate ($n = 500$), and large ($n = 750$) local sample sizes. The selected local sample sizes are adapted from sample size requirements for ETS DIF analysis (Zwick, 2012) and from suggested IRT model calibration sample sizes (de Ayala, 2009, p. 105). To observe the effect of a mixture of sample sizes, which is likely to occur in practice, local samples sizes of $n = 250, 500, and 750$ will be allocated in accordance with the relative size of the 2016-2017 school enrollments in each of the 75 Arkansas counties (Arkansas Department of Education [ADE], 2017). The manipulation of this factor will show the impact of different sample sizes on the proposed methodology in practice.

Other factors that will be manipulated in the case studies include the size of the bandwidth, which will be referred to as bandwidth selection, and the local IRT model type. Despite the bandwidths being held fixed (rather than adaptive), three fixed values will be utilized in each simulated case study to demonstrate the sensitivity of local IRT results to bandwidth selection. Furthermore, changes in the bandwidth selection will demonstrate various levels of smoothness in the surface of parameter estimates. Both 1PL and 2PL local IRT models will be utilized in each simulated case study to investigate the potential of the proposed method for detecting regional DIF. The 1PL local IRT model will be used to test for uniform regional DIF while the more flexible 2PL local IRT model will be used to test for both uniform and non-uniform regional DIF.

All of the manipulated factors in the demonstrative case studies utilizing simulated data are selected to reveal local IRT model sensitivities to spatial structures, types of DIF, magnitudes of DIF, bandwidth selections, local sample sizes, and IRT model choices. The information gathered can be used to guide the future development and use of the proposed method in practice.
Empirical Data

The proposed methodology will also be demonstrated with a case study application to empirical survey data. These data and the instrument items to be investigated for regional DIF are described below.

Data source and sample. Data from the 2015-16 Malawi Demographic and Health Survey (DHS) provide an empirical application of the currently proposed method for regional DIF detection. As of June 2017, these data represent the most recent DHS survey data collected in Malawi. According to the Malawi National Statistical Office (NSO), the sampling frame used for the 2015-16 DHS, consisting of stratification and proportional allocation prior to a multiple stage selection process, is designed to be representative “for the country as a whole, for urban and rural areas separately, and for each of the 28 districts” (NSO, 2017, p. 2).

While the 2015-16 Malawi DHS consisted of four separate questionnaires, only the woman’s questionnaire will be utilized for the purposes of the current study. The woman’s questionnaire collects information on all eligible women in the sample (age 15-49 years) regarding a variety of topics including but not limited to fertility, family planning, marriage, and domestic violence. However, the focus of this empirical case study will be on a portion of the woman’s questionnaire that assesses family planning method (i.e., contraception) knowledge. According to the NSO, the woman’s questionnaire in Malawi had a response rate of 98% with 24,562 of the 25,146 eligible women successfully completing interviews (NSO, 2017, p. 7).

Though 98% of women in Malawi have a knowledge of contraceptive methods, defined as knowing at least one method of contraception, 37% of those using a contraceptive method discontinue the method in less than 12 months. There appears to be a substantial unmet need for family planning with 41% of all births in the past five years not being wanted at the time of
conception, exposure to family planning messages in the media is still limited with 42% of women having no exposure to these messages in recent months, and there may exist a gap in family planning knowledge due to differences in an interaction of factors related to wealth, education, geographic location, and the availability of fieldworkers or healthcare facilities to discuss family planning methods (NSO, 2017). While women may know one contraceptive method, greater knowledge of contraceptive methods would allow for better health-related contraceptive choices, alternative contraceptive choices following discontinuation of a method, and more ability to meet the contraceptive needs of Malawian women thereby limiting the risks of either childbirth or abortion services when access to adequate medical care is not available. Consequently, there are also motivating reasons for healthcare professionals to be concerned with regional differences not only in the access to contraceptive methods but also in the knowledge of particular contraceptive methods that might more adequately meet the need for family planning in Malawi.

**Instrument and item format.** The instrument to be investigated for regional DIF is a 13 item dichotomously scored (0 = “No”, 1 = “Yes”) instrument intended to measure an individual’s level of family planning method knowledge obtained from the most recent DHS survey data collected in Malawi. The instrument items are provided in Table 3.

It is quite possible that these survey items might function differently across space in Malawi due to regional disparities and inequities. For this reason, the proposed localized approach to IRT for regional DIF detection will be applied to this 13 item instrument extracted from the 2015-16 Malawi DHS data. This information can be of benefit to global health organizations, healthcare professionals, and women in Malawi that desire increased family
planning method knowledge with the goal of increasing the health of all women, children and families in Malawi regardless of geographic location of residence.

Table 3. The Malawi DHS complete family planning knowledge instrument items.

<table>
<thead>
<tr>
<th>Item</th>
<th>Item Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>V304$01 (Q1)</td>
<td>Knows Method - Pill</td>
</tr>
<tr>
<td>V304$02 (Q2)</td>
<td>Knows Method - IUD</td>
</tr>
<tr>
<td>V304$03 (Q3)</td>
<td>Knows Method - Injections</td>
</tr>
<tr>
<td>V304$05 (Q4)</td>
<td>Knows Method - Male Condom</td>
</tr>
<tr>
<td>V304$06 (Q5)</td>
<td>Knows Method - Female Sterilization</td>
</tr>
<tr>
<td>V304$07 (Q6)</td>
<td>Knows Method - Male Sterilization</td>
</tr>
<tr>
<td>V304$08 (Q7)</td>
<td>Knows Method - Periodic Abstinence</td>
</tr>
<tr>
<td>V304$09 (Q8)</td>
<td>Knows Method - Withdrawal</td>
</tr>
<tr>
<td>V304$11 (Q9)</td>
<td>Knows Method - Implants</td>
</tr>
<tr>
<td>V304$13 (Q10)</td>
<td>Knows Method - Lactational Amenorrhea (LAM) Method</td>
</tr>
<tr>
<td>V304$14 (Q11)</td>
<td>Knows Method - Female Condom</td>
</tr>
<tr>
<td>V304$16 (Q12)</td>
<td>Knows Method - Emergency Contraception</td>
</tr>
<tr>
<td>V304$18 (Q13)</td>
<td>Knows Method - Standard Days Method (SDM)</td>
</tr>
</tbody>
</table>

Case Study Descriptions

Case studies, in the form of applications to both simulated and empirical data, will demonstrate the utility and the benefit of localized IRT modeling as an exploratory tool for regional DIF detection, especially when comparisons are made to traditional, non-proximal DIF techniques. There is an illustrative purpose for each case study setting described below.

The first three case studies (CS 1, 2, and 3) will be set on an irregular lattice with 75 areal units as seen in Figure 17 while the last case study (CS 4) is an application of the proposed methodology to the most recent DHS survey data collected in Malawi (as of June 2017) and, as such, will be set on an irregular lattice with 28 areal units presented in CS 4. Note that the manipulated factors for all case study settings are summarized and can be compared in Table 4.
Case study one. The first case study includes only one latent spatial subgroup so that all 75 areal units are simulated to have the same true item parameter values and ability distributions. This case study, being set on an irregular lattice, is placed in the context of a realistic spatial structure where additional complexities in local modeling emerge. Local sample size considerations are, now, potentially more important and have a greater impact on results than would be the case on a regular lattice. Relatedly, the use of a fixed bandwidth given unequal local calibration sample sizes can affect results. Data is simulated for the 75 counties of Arkansas shown in Figure 17. The centroids of each county are also provided in Figure 17, as these will serve as the regression points for local model calibration. All of the 75 areal units are simulated to have the same true item parameter values and ability distributions. Consequently, no regional DIF should be detected by the proposed method, making the overwhelming \( \binom{75}{2} = 2775 \) pairwise comparisons typical in traditional DIF techniques unnecessary. The primary aim of this

![Arkansas Counties Centroids](image)

*Figure 17. Irregular lattice of Arkansas counties for CS 1, 2, and 3.*
case study will be to investigate the impact of bandwidth selection, local sample size variation, and model type on the proposed method results. Owing to the potential for inflated type I error rates in local modeling, the sensitivity of the method to sampling variability will be demonstrated as several factors are manipulated. This case study will also provide the reader with an initial visual example of spatial stationarity and an introductory demonstration of the smoothing effect of bandwidth selections. This first case study involves an irregular lattice where local calibration sample sizes will differ due, in large part, to the use of a fixed bandwidth with areal units that differ in both geographic size and, also, in sample size. Consequently, while areal unit sample size will be manipulated in the current case study, the assignment of the different local sample sizes, \( n = 250, 500, 750 \), will be allocated in accordance with the relative size of the 2016-2017 school enrollments in each county as seen in Figure 18 (ADE, 2017). Those counties with school

![Arkansas Counties 2016-2017 School Enrollment](image)

*Figure 18. Arkansas county relative enrollment sizes.*
enrollments in the lower quartile will be assigned sample sizes of $n = 250$ whereas counties with school enrollments between the first and third quartiles will be assigned sample sizes of $n = 500$ and counties with school enrollments in the upper quartile will be assigned sample sizes of $n = 750$. It is worth noting that the simulated sample sizes are realistic given the county school enrollment numbers.

The sensitivity of the proposed method will be demonstrated as several factors are manipulated. This case study will also begin to investigate MAUP issues in a more realistic context, though these issues will be more thoroughly investigated in the later case studies.

**Case study two.** The second case study includes two latent spatial subgroups among the 75 Arkansas counties displayed in Figure 17. The areal units will be assigned to two latent spatial subgroups based upon their inclusion or lack thereof in either of two metropolitan statistical areas (MSAs) in the state that surround two major universities as seen in Figure 19.

![Arkansas Counties Metropolitan Statistical Areas](image)

**Figure 19.** Arkansas metropolitan statistical areas for CS 2.
Consequently, counties that are a part of the Fayetteville-Springdale-Rogers MSA (i.e., Benton, Madison, and Washington counties) or are a part of the Little Rock-North Little Rock-Conway MSA (i.e., Faulkner, Grant, Lonoke, Perry, Pulaski, and Saline counties) will serve as one latent spatial subgroup while all of the remaining Arkansas counties will serve as the second latent spatial subgroup. All of the county local sample size allocations will be made according to the 2016-2017 county enrollment numbers as described in case study one. The counties belonging to the MSAs seen in Figure 19, will be simulated to have uniform and, also, non-uniform DIF. The magnitude of the DIF will be manipulated such that an areal unit in one of the latent subgroups, when compared to an areal unit in the other latent subgroup, would have an area between their respective IRFs of 0.40, 0.60, or 0.80. Consequently, regional DIF should be detected by the proposed method and two latent spatial subgroups should be identified greatly simplifying subsequent, confirmatory DIF testing.

However, the primary aim of this second case study will be to demonstrate the benefit of the proposed method when comparisons are made to that of traditional DIF testing using a non-proximal approach with spatial groups defined using cardinal direction as in Figure 20. As previously suggested, this type of latent spatial structure might render traditional DIF techniques comparing regions North to South or East to West inadequate. Further, these traditional DIF techniques are prone to a spatial Simpson’s paradox where this induced regional DIF would be potentially unobservable on a more aggregate, global level.

**Case study three.** The final simulated case study includes an unknown number of latent spatial subgroups. The reason that this case study is described as having an unknown number of spatial subgroups is due to the fact that the regions exhibiting DIF will be randomly assigned
across the spatial study region, which may or may not create discernable spatial subgroups (i.e., clusters). Data will be simulated for all of the 75 counties by simulating 50 reference samples and 25 focal samples, which exhibit a certain type and magnitude of DIF. Consequently, there will exist areal units that exhibit DIF but this DIF may or may not have a spatial structure. Once certain areal units (i.e., counties) have been randomly assigned to be either a reference or a focal group member, sample size allocations will be made according to the 2016-2017 county enrollment numbers as described in case studies one and two. The data for each county can then be simulated using WinGen3 (Han, 2007). Both uniform and non-uniform DIF will be simulated for the focal group samples with magnitudes manipulated such that the area between IRFs for a reference group areal unit and a focal group areal unit will be 0.40, 0.60, or 0.80.
This simulated case study is the most similar to an empirical study since regional DIF with identified spatial subgroups may or may not be present. DIF is present in one third of the counties but may not be present in a form that displays spatial structure. Rather, it might be present in a form where multiple group DIF detection techniques or many pairwise comparisons would be more reasonable if the number of subgroups present is small. The proposed method will be employed with the goal of exploring potential regional DIF and spatial nonstationarity that could be beneficial when a large number of subgroups exist. An attempt will be made to detect regional DIF and to identify spatial subgroups for subsequent, confirmatory DIF testing. This case study analysis will be carried out under the manipulation of several circumstantial factors that cannot be controlled empirically such as DIF type, DIF magnitude, and local sample size. However, the analysis will provide a detailed overview of the manipulation of user-specified factors such as bandwidth selection and local IRT model type. Comparisons will be made between the results of the proposed methodology and of traditional DIF techniques utilizing a non-proximal approach. The non-proximal approach refers to the point at which the study region is split into groups based upon cardinal direction only and comparisons are made between North and South or East and West regions, as seen in Figure 20.

Case study four. This fourth case study is an empirical application of the currently proposed method for regional DIF detection to a 13 item instrument intended to measure the level of family planning method knowledge, which is extracted from the most recent DHS survey data collected in Malawi (as of June 2017). This empirical data is set then on an irregular lattice with 28 areal units as seen in Figure 21. The centroids of each county are also provided in Figure 21, as these will serve as the regression points for local model calibration.
Though GPS latitude/longitude positions for each responding household are collected for the Malawi DHS survey, these positions are randomly displaced to ensure respondent confidentiality. Urban locations are displaced up to 2 kilometers while rural locations are displaced up to 5 kilometers (with 1% of rural locations displaced up to 10 kilometers). Despite this random displacement of GPS latitude/longitude positions whereby locations are geographically off-set for respondent confidentiality, the displacement is restricted to the second administrative level boundaries in a country. Consequently, displaced/off-set locations remain within the same country, region, and district as the undisplaced/original locations. Given the inherent error in these geographic coordinates but the confinement of any geographic
displacement of coordinates to the district level, it is reasonable to aggregate data to the level of the 28 areal units representing Malawi districts. This geographic off-setting of GPS data in the DHS program is the reasoning behind the regression points being allocated to the 28 Malawi district centroids seen in Figure 21.

While the third simulated case study will have some similarities to an empirical study, the application of the proposed methodology to actual survey data, especially on an irregular lattice with a more unusual shape than that of Arkansas, will present additional challenges. The proposed method will again be employed using four different fixed, user-specified bandwidths with the goal of exploring potential regional DIF and spatial nonstationarity that exists in the survey items intended to measure the knowledge of family planning methods.

These survey items might function differently across space in Malawi due to regional disparities and inequities. While observable regional disparities in item functionality might be directly attributable to differences in the spatial location of observations, any detectable differences in item functionality also highlight the need for further investigation into alternate explanatory covariates, such as specific area social and cultural beliefs and practices. This information can be invaluable to healthcare professionals, striving to increase the family planning method knowledge of Malawian women regardless of geographic location of residence.

The application of this localized approach to IRT for DIF detection may assist in equalizing family planning method opportunity and knowledge for all women in Malawi no matter their geographic location, while also providing a detailed overview of the method in context. Comparisons will again be made between the results of the proposed methodology and of traditional DIF techniques utilizing a non-proximal approach based upon the pre-specified Malawian regions (Northern, Central, Southern) as seen in Figure 22.
Figure 22. Malawi district subgroups defined by cardinal direction.

The proposed method serves to detect DIF across space without a priori groupings, identifying regional disparities and latent spatial trends that may otherwise be unobservable. The three simulated case studies and the one empirical application will serve as detailed examples to illustrate the use of the proposed method, while investigating potential local IRT model sensitivities to spatial structures, types of DIF, magnitudes of DIF, bandwidth selections, local sample sizes, and IRT model choices. Each individual case study provided attempts to demonstrate hypothesized benefits while also exploring possible disadvantages of the proposed method given various factor modifications.
Table 4. Case study descriptions of non-manipulated and manipulated factors.

<table>
<thead>
<tr>
<th>Case Study</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4 - Empirical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent Spatial Groups</td>
<td>One</td>
<td>Two</td>
<td>Two, Randomly Distributed Across the Spatial Region</td>
<td>Unknown</td>
</tr>
<tr>
<td>DIF Type</td>
<td>N/A – No DIF</td>
<td>Uniform, Non-Uniform</td>
<td>Uniform, Non-Uniform</td>
<td>Unknown</td>
</tr>
<tr>
<td>DIF Magnitude</td>
<td>N/A – No DIF</td>
<td>Area between IRFs: 0.40, 0.60, 0.80</td>
<td>Area between IRFs: 0.40, 0.60, 0.80</td>
<td>Area between IRFs: Unknown</td>
</tr>
<tr>
<td>Ability Difference</td>
<td>No Difference</td>
<td>No Difference</td>
<td>No Difference</td>
<td>Unknown</td>
</tr>
<tr>
<td>Lattice Type</td>
<td>Irregular, 75 Areal Units</td>
<td>Irregular, 75 Areal Units</td>
<td>Irregular, 75 Areal Units</td>
<td>Irregular, 28 Areal Units</td>
</tr>
<tr>
<td>Bandwidth Type</td>
<td>Fixed</td>
<td>Fixed</td>
<td>Fixed</td>
<td>Fixed</td>
</tr>
<tr>
<td>Bandwidth Selection</td>
<td>$h = 25$ km, $h = 50$ km, $h = 75$ km</td>
<td>$h = 25$ km, $h = 50$ km, $h = 75$ km</td>
<td>$h = 25$ km, $h = 50$ km, $h = 75$ km</td>
<td>$h = 20$ km, $h = 80$ km, $h = 140$ km, $h = 200$ km</td>
</tr>
<tr>
<td>Sample Size</td>
<td>Total Sample Size $N = 37500$, 3 Local Sizes $n = 250,500,750^*$</td>
<td>Total Sample Size $N = 37500$, 3 Local Sizes $n = 250,500,750^*$</td>
<td>Total Sample Size $N = 37500$, 3 Local Sizes $n = 250,500,750^*$</td>
<td>Total Sample Size $N = 24562$, Local Sizes Vary</td>
</tr>
<tr>
<td>Local IRT Model Type</td>
<td>1PL, 2PL</td>
<td>1PL, 2PL</td>
<td>1PL, 2PL</td>
<td>1PL and/or 2PL</td>
</tr>
</tbody>
</table>

* allocated in accordance with the relative size of the 2016-2017 county school enrollments
Chapter 4

Results

Divided into four major sections corresponding to the four case studies described previously, the current chapter presents the relevant results of the aforementioned case studies as well as providing additional discussion of these results. Each of the four case studies, utilizing either simulated or empirical data, is intended to illustrate through demonstration the use of a localized approach to IRT modeling for the detection of item-level regional DIF based upon geographic location without the need for pre-specified groupings and before any confirmatory DIF testing. Within these illustrative examples of the proposed method, both advantages and disadvantages will be addressed. Additionally, several procedural choices such as bandwidth selection and model choice (i.e., 1PL or 2PL) in the presence of different local sample sizes, DIF types, DIF magnitudes, and spatial structures will be discussed.

The results of applying a localized approach to IRT modeling in the context of each case study setting will be presented first for the 1PL local IRT model and then for the 2PL local IRT model. Following the presentation of relevant method results within each case study, a brief description of the primary observations and concepts of interest will be provided.

Case Study 1 – Simulated Data Exhibiting No Regional DIF

The first three case studies apply the proposed localized approach to IRT modeling to simulated data. While the simulation of data for the 75 counties in the state of Arkansas was discussed previously, it is of note that data were simulated separately for each county in the state of Arkansas assuming equal ability distributions such that \( \theta \sim N(0, 1) \).

In this first case study, two sets of data were simulated corresponding to either a 1PL or a 2PL IRT model. For the 1PL simulated data, true difficulty parameters for items were drawn
from a $N(0, 1)$ distribution while the true discrimination parameter for items was constrained to be the unit discrimination parameter characteristic of Rasch models. Response strings for examinees in each of the 75 Arkansas counties were then simulated based upon these randomly selected true parameter values. For the 2PL simulated data, true difficulty parameters for items were drawn from a $N(0, 1)$ distribution and true discrimination parameters for items were drawn from a $\text{Unif}(0.25, 1.75)$ distribution. As in the 1PL setting, response strings for examinees in each of the 75 Arkansas counties were then simulated based upon these randomly selected true parameter values. The number of examinees in each county was allocated in accordance with the relative size of the 2016-2017 school enrollments as discussed previously and as seen in Figure 18. Local sample size allocations for each county are provided in Table 5.

Table 5. Assigned county sample sizes.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Counties</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n=250$</td>
<td>Calhoun, Chicot, Cleveland, Dallas, Fulton, Lafayette, Lee, Lincoln, Marion, Monroe, Montgomery, Nevada, Newton, Perry, Prairie, Scott, Searcy, Stone, Woodruff</td>
</tr>
<tr>
<td>$n=500$</td>
<td>Arkansas, Ashley, Baxter, Boone, Bradley, Carroll, Clark, Clay, Cleburne, Columbia, Conway, Cross, Desha, Drew, Franklin, Grant, Hempstead, Hot Spring, Howard, Izard, Jackson, Johnson, Lawrence, Little River, Logan, Madison, Ouachita, Phillips, Pike, Poinsett, Polk, Randolph, St. Francis, Sevier, Sharp, Van Buren, Yell</td>
</tr>
<tr>
<td>$n=750$</td>
<td>Benton, Craighead, Crawford, Crittenden, Faulkner, Garland, Greene, Independence, Jefferson, Lonoke, Miller, Mississippi, Pope, Pulaski, Saline, Sebastian, Union, Washington, White</td>
</tr>
</tbody>
</table>

The centroids of each county serve as the regression points for local model calibration in both the 1PL and 2PL settings. These centroids were previously displayed in Figure 17. The proposed method was employed in each of the simulated case studies using three different fixed,
user-specified bandwidths \((h = 25\text{km}, 50\text{km}, 75\text{km})\) with the goal of exploring potential regional DIF and spatial nonstationarity that exists in the items. The neighborhood structures that are formed by these three specified bandwidths can be seen in Figure 23. It is of note that the smallest bandwidth \((h = 25\text{km})\) is equivalent to an areal unit calibration of the global model, a precursor to moving window and geographically weighted local spatial modeling techniques. The 50 km bandwidth, however, results in a neighborhood structure that has an average of 3.12 links/neighbors for each calibration point and at most six links. The 75 km bandwidth expands the spatial neighborhoods, having an average of 7.09 links/neighbors and as many as ten links.

![Figure 23. Neighborhood structures for Arkansas counties at three fixed bandwidths.](image)

The primary aim of this first case study was to investigate the impact of bandwidth selection, local (i.e., calibration) sample size, and model type on the proposed method results when only one latent spatial subgroup is present and, as such, no regional DIF should be detected. The results will be presented for the 1PL local IRT model and the 2PL local IRT model before discussing the core objectives and findings of the case study.

**1PL local IRT model results.** The localized approach to IRT modeling and DIF detection, utilizing a box-car kernel weighting scheme with three fixed bandwidths, was applied
to the Arkansas county 25-item instrument data, simulated to exhibit no regional DIF. Table 6 provides the global 1PL IRT difficulty estimates and associated standard errors for all 25 items as well as providing the 1PL IRT difficulty estimates and associated standard errors for all 25 items averaged across the 75 local calibrations of the 1PL model for each of the three different fixed bandwidths. For all 1PL model calibrations (including the global model that incorporated all simulated instrument response strings for the 37,500 examinees statewide), the discrimination parameter was constrained to one and, consequently, is excluded from Table 6. Figure 24 and Tables 7 - 8 provide county-specific difficulty estimates and associated standard errors for the local calibrations of the 1PL IRT models at each of three different bandwidths for Question 1.

Table 6. 1PL local IRT model summary – CS 1 estimated difficulty coefficients averaged across counties.

<table>
<thead>
<tr>
<th>Item (Question)</th>
<th>Global Model Estimated Difficulty</th>
<th>Global Model Standard Error</th>
<th>25 km Mean Estimated Difficulty</th>
<th>25 km Mean Standard Error</th>
<th>50 km Mean Estimated Difficulty</th>
<th>50 km Mean Standard Error</th>
<th>75 km Mean Estimated Difficulty</th>
<th>75 km Mean Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.027</td>
<td>0.013</td>
<td>-0.027</td>
<td>0.115</td>
<td>-0.017</td>
<td>0.056</td>
<td>-0.025</td>
<td>0.039</td>
</tr>
<tr>
<td>2</td>
<td>0.929</td>
<td>0.013</td>
<td>0.926</td>
<td>0.121</td>
<td>0.924</td>
<td>0.060</td>
<td>0.918</td>
<td>0.042</td>
</tr>
<tr>
<td>3</td>
<td>-0.308</td>
<td>0.013</td>
<td>-0.308</td>
<td>0.115</td>
<td>-0.309</td>
<td>0.057</td>
<td>-0.307</td>
<td>0.040</td>
</tr>
<tr>
<td>4</td>
<td>1.648</td>
<td>0.015</td>
<td>1.654</td>
<td>0.137</td>
<td>1.642</td>
<td>0.067</td>
<td>1.643</td>
<td>0.047</td>
</tr>
<tr>
<td>5</td>
<td>-0.343</td>
<td>0.013</td>
<td>-0.341</td>
<td>0.115</td>
<td>-0.342</td>
<td>0.057</td>
<td>-0.341</td>
<td>0.040</td>
</tr>
<tr>
<td>6</td>
<td>-1.344</td>
<td>0.014</td>
<td>-1.350</td>
<td>0.129</td>
<td>-1.340</td>
<td>0.063</td>
<td>-1.346</td>
<td>0.044</td>
</tr>
<tr>
<td>7</td>
<td>-0.638</td>
<td>0.013</td>
<td>-0.641</td>
<td>0.118</td>
<td>-0.634</td>
<td>0.058</td>
<td>-0.639</td>
<td>0.040</td>
</tr>
<tr>
<td>8</td>
<td>1.431</td>
<td>0.014</td>
<td>1.427</td>
<td>0.130</td>
<td>1.432</td>
<td>0.064</td>
<td>1.428</td>
<td>0.045</td>
</tr>
<tr>
<td>9</td>
<td>1.627</td>
<td>0.015</td>
<td>1.631</td>
<td>0.136</td>
<td>1.627</td>
<td>0.067</td>
<td>1.622</td>
<td>0.047</td>
</tr>
<tr>
<td>10</td>
<td>-0.032</td>
<td>0.013</td>
<td>-0.031</td>
<td>0.115</td>
<td>-0.029</td>
<td>0.056</td>
<td>-0.034</td>
<td>0.039</td>
</tr>
<tr>
<td>11</td>
<td>-1.509</td>
<td>0.014</td>
<td>-1.514</td>
<td>0.133</td>
<td>-1.503</td>
<td>0.065</td>
<td>-1.505</td>
<td>0.046</td>
</tr>
<tr>
<td>12</td>
<td>-0.099</td>
<td>0.013</td>
<td>-0.099</td>
<td>0.115</td>
<td>-0.096</td>
<td>0.056</td>
<td>-0.101</td>
<td>0.039</td>
</tr>
<tr>
<td>13</td>
<td>1.711</td>
<td>0.015</td>
<td>1.720</td>
<td>0.139</td>
<td>1.710</td>
<td>0.068</td>
<td>1.705</td>
<td>0.047</td>
</tr>
<tr>
<td>14</td>
<td>-0.227</td>
<td>0.013</td>
<td>-0.223</td>
<td>0.115</td>
<td>-0.228</td>
<td>0.057</td>
<td>-0.231</td>
<td>0.040</td>
</tr>
<tr>
<td>15</td>
<td>-0.355</td>
<td>0.013</td>
<td>-0.352</td>
<td>0.116</td>
<td>-0.361</td>
<td>0.057</td>
<td>-0.356</td>
<td>0.040</td>
</tr>
<tr>
<td>16</td>
<td>-0.729</td>
<td>0.013</td>
<td>-0.722</td>
<td>0.118</td>
<td>-0.733</td>
<td>0.058</td>
<td>-0.729</td>
<td>0.041</td>
</tr>
<tr>
<td>17</td>
<td>0.923</td>
<td>0.013</td>
<td>0.935</td>
<td>0.121</td>
<td>0.927</td>
<td>0.060</td>
<td>0.923</td>
<td>0.042</td>
</tr>
<tr>
<td>18</td>
<td>2.777</td>
<td>0.020</td>
<td>2.790</td>
<td>0.187</td>
<td>2.767</td>
<td>0.091</td>
<td>2.772</td>
<td>0.064</td>
</tr>
<tr>
<td>19</td>
<td>-0.202</td>
<td>0.013</td>
<td>-0.204</td>
<td>0.115</td>
<td>-0.197</td>
<td>0.056</td>
<td>-0.203</td>
<td>0.040</td>
</tr>
<tr>
<td>20</td>
<td>-0.773</td>
<td>0.013</td>
<td>-0.772</td>
<td>0.119</td>
<td>-0.771</td>
<td>0.059</td>
<td>-0.770</td>
<td>0.041</td>
</tr>
<tr>
<td>21</td>
<td>1.396</td>
<td>0.014</td>
<td>1.392</td>
<td>0.130</td>
<td>1.396</td>
<td>0.064</td>
<td>1.392</td>
<td>0.045</td>
</tr>
<tr>
<td>22</td>
<td>-1.910</td>
<td>0.016</td>
<td>-1.909</td>
<td>0.144</td>
<td>-1.910</td>
<td>0.071</td>
<td>-1.907</td>
<td>0.050</td>
</tr>
<tr>
<td>23</td>
<td>0.769</td>
<td>0.013</td>
<td>0.767</td>
<td>0.119</td>
<td>0.768</td>
<td>0.059</td>
<td>0.766</td>
<td>0.041</td>
</tr>
<tr>
<td>24</td>
<td>-0.606</td>
<td>0.013</td>
<td>-0.607</td>
<td>0.117</td>
<td>-0.601</td>
<td>0.058</td>
<td>-0.605</td>
<td>0.040</td>
</tr>
<tr>
<td>25</td>
<td>-2.089</td>
<td>0.017</td>
<td>-2.086</td>
<td>0.151</td>
<td>-2.089</td>
<td>0.074</td>
<td>-2.089</td>
<td>0.052</td>
</tr>
</tbody>
</table>
Figure 24. 1PL local IRT model difficulty estimates and standard errors for CS 1.
Table 7. *IPL local IRT model summary* – *CS 1 estimated difficulty coefficients for question 1 by county.*

<table>
<thead>
<tr>
<th>County</th>
<th>25 km Estimated Difficulty</th>
<th>50 km Estimated Difficulty</th>
<th>75 km Estimated Difficulty</th>
<th>County</th>
<th>25 km Estimated Difficulty</th>
<th>50 km Estimated Difficulty</th>
<th>75 km Estimated Difficulty</th>
<th>County</th>
<th>25 km Estimated Difficulty</th>
<th>50 km Estimated Difficulty</th>
<th>75 km Estimated Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arkansas</td>
<td>-0.080</td>
<td>-0.016</td>
<td>-0.010</td>
<td>Garland</td>
<td>-0.023</td>
<td>-0.020</td>
<td>0.000</td>
<td>Newton</td>
<td>-0.155</td>
<td>-0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>Ashley</td>
<td>0.049</td>
<td>0.070</td>
<td>0.070</td>
<td>Grant</td>
<td>0.007</td>
<td>-0.026</td>
<td>-0.012</td>
<td>Ouachita</td>
<td>0.002</td>
<td>-0.066</td>
<td>-0.018</td>
</tr>
<tr>
<td>Baxter</td>
<td>0.089</td>
<td>0.023</td>
<td>-0.045</td>
<td>Greene</td>
<td>-0.051</td>
<td>-0.050</td>
<td>-0.004</td>
<td>Perry</td>
<td>0.264</td>
<td>-0.014</td>
<td>-0.004</td>
</tr>
<tr>
<td>Benton</td>
<td>-0.055</td>
<td>-0.050</td>
<td>-0.026</td>
<td>Hempstead</td>
<td>-0.019</td>
<td>0.006</td>
<td>-0.033</td>
<td>Phillips</td>
<td>-0.022</td>
<td>-0.027</td>
<td>-0.058</td>
</tr>
<tr>
<td>Boone</td>
<td>-0.080</td>
<td>-0.081</td>
<td>-0.011</td>
<td>Hot Spring</td>
<td>-0.058</td>
<td>-0.044</td>
<td>-0.015</td>
<td>Pike</td>
<td>0.037</td>
<td>0.019</td>
<td>0.047</td>
</tr>
<tr>
<td>Bradley</td>
<td>0.196</td>
<td>0.098</td>
<td>0.008</td>
<td>Howard</td>
<td>-0.016</td>
<td>-0.083</td>
<td>-0.079</td>
<td>Poinsett</td>
<td>0.056</td>
<td>-0.035</td>
<td>0.007</td>
</tr>
<tr>
<td>Calhoun</td>
<td>-0.037</td>
<td>-0.012</td>
<td>-0.000</td>
<td>Independence</td>
<td>-0.051</td>
<td>-0.035</td>
<td>-0.042</td>
<td>Polk</td>
<td>-0.262</td>
<td>-0.159</td>
<td>-0.082</td>
</tr>
<tr>
<td>Carroll</td>
<td>-0.166</td>
<td>-0.019</td>
<td>-0.044</td>
<td>Izard</td>
<td>-0.061</td>
<td>-0.040</td>
<td>-0.012</td>
<td>Pope</td>
<td>0.078</td>
<td>0.036</td>
<td>0.036</td>
</tr>
<tr>
<td>Chicot</td>
<td>-0.120</td>
<td>-0.007</td>
<td>0.002</td>
<td>Jackson</td>
<td>0.003</td>
<td>0.001</td>
<td>-0.028</td>
<td>Prairie</td>
<td>0.351</td>
<td>0.017</td>
<td>0.015</td>
</tr>
<tr>
<td>Clay</td>
<td>-0.041</td>
<td>-0.047</td>
<td>-0.012</td>
<td>Johnson</td>
<td>0.062</td>
<td>-0.024</td>
<td>-0.007</td>
<td>Randolph</td>
<td>-0.093</td>
<td>-0.001</td>
<td>0.017</td>
</tr>
<tr>
<td>Clarks</td>
<td>-0.040</td>
<td>-0.102</td>
<td>-0.090</td>
<td>Lafayette</td>
<td>-0.194</td>
<td>-0.076</td>
<td>-0.051</td>
<td>St. Francis</td>
<td>-0.099</td>
<td>-0.024</td>
<td>-0.006</td>
</tr>
<tr>
<td>Cleveland</td>
<td>0.213</td>
<td>0.039</td>
<td>0.037</td>
<td>Lawrence</td>
<td>0.178</td>
<td>0.007</td>
<td>-0.017</td>
<td>Saline</td>
<td>-0.084</td>
<td>0.014</td>
<td>0.033</td>
</tr>
<tr>
<td>Columbia</td>
<td>-0.174</td>
<td>-0.000</td>
<td>-0.058</td>
<td>Lee</td>
<td>-0.135</td>
<td>-0.050</td>
<td>0.013</td>
<td>Scott</td>
<td>-0.242</td>
<td>-0.092</td>
<td>-0.084</td>
</tr>
<tr>
<td>Conway</td>
<td>-0.044</td>
<td>0.001</td>
<td>-0.009</td>
<td>Lincoln</td>
<td>0.019</td>
<td>0.025</td>
<td>0.046</td>
<td>Searcy</td>
<td>-0.197</td>
<td>-0.075</td>
<td>-0.021</td>
</tr>
<tr>
<td>Craighead</td>
<td>-0.027</td>
<td>0.023</td>
<td>-0.177</td>
<td>Little River</td>
<td>-0.199</td>
<td>-0.085</td>
<td>-0.030</td>
<td>Sebastian</td>
<td>0.070</td>
<td>-0.031</td>
<td>0.061</td>
</tr>
<tr>
<td>Crawford</td>
<td>-0.065</td>
<td>-0.016</td>
<td>-0.004</td>
<td>Logan</td>
<td>-0.198</td>
<td>-0.050</td>
<td>-0.022</td>
<td>Sevier</td>
<td>-0.041</td>
<td>-0.085</td>
<td>0.083</td>
</tr>
<tr>
<td>Crittenden</td>
<td>0.080</td>
<td>-0.034</td>
<td>-0.023</td>
<td>Lonoke</td>
<td>-0.161</td>
<td>0.036</td>
<td>-0.039</td>
<td>Sharp</td>
<td>-0.013</td>
<td>-0.018</td>
<td>-0.030</td>
</tr>
<tr>
<td>Cross</td>
<td>-0.140</td>
<td>0.003</td>
<td>-0.011</td>
<td>Madison</td>
<td>0.190</td>
<td>-0.030</td>
<td>-0.035</td>
<td>Stone</td>
<td>-0.231</td>
<td>0.083</td>
<td>0.035</td>
</tr>
<tr>
<td>Dallas</td>
<td>-0.338</td>
<td>-0.007</td>
<td>0.031</td>
<td>Marion</td>
<td>0.158</td>
<td>-0.003</td>
<td>-0.052</td>
<td>Union</td>
<td>-0.121</td>
<td>-0.099</td>
<td>-0.035</td>
</tr>
<tr>
<td>Desha</td>
<td>-0.040</td>
<td>-0.021</td>
<td>-0.028</td>
<td>Miller</td>
<td>0.029</td>
<td>-0.028</td>
<td>-0.078</td>
<td>Van Buren</td>
<td>-0.017</td>
<td>-0.028</td>
<td>0.061</td>
</tr>
<tr>
<td>Drew</td>
<td>0.059</td>
<td>0.090</td>
<td>0.051</td>
<td>Mississippi</td>
<td>-0.012</td>
<td>-0.012</td>
<td>0.000</td>
<td>Washington</td>
<td>-0.045</td>
<td>-0.011</td>
<td>-0.033</td>
</tr>
<tr>
<td>Faulkner</td>
<td>-0.149</td>
<td>-0.008</td>
<td>-0.045</td>
<td>Monroe</td>
<td>0.074</td>
<td>0.012</td>
<td>-0.034</td>
<td>White</td>
<td>-0.207</td>
<td>-0.091</td>
<td>-0.042</td>
</tr>
<tr>
<td>Franklin</td>
<td>-0.028</td>
<td>-0.025</td>
<td>-0.018</td>
<td>Montgomery</td>
<td>-0.017</td>
<td>-0.001</td>
<td>-0.048</td>
<td>Woodruff</td>
<td>0.159</td>
<td>-0.053</td>
<td>-0.039</td>
</tr>
<tr>
<td>Fulton</td>
<td>-0.082</td>
<td>-0.008</td>
<td>-0.018</td>
<td>Nevada</td>
<td>0.035</td>
<td>0.026</td>
<td>-0.033</td>
<td>Yell</td>
<td>-0.036</td>
<td>-0.042</td>
<td>-0.013</td>
</tr>
</tbody>
</table>
Table 8. IPL local IRT model summary – CS 1 standard errors for estimated difficulty coefficients for question 1 by county.

<table>
<thead>
<tr>
<th>County</th>
<th>25 km Estimated Difficulty</th>
<th>50 km Estimated Difficulty</th>
<th>75 km Estimated Difficulty</th>
<th>County</th>
<th>25 km Estimated Difficulty</th>
<th>50 km Estimated Difficulty</th>
<th>75 km Estimated Difficulty</th>
<th>County</th>
<th>25 km Estimated Difficulty</th>
<th>50 km Estimated Difficulty</th>
<th>75 km Estimated Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arkansas</td>
<td>0.108</td>
<td>0.077</td>
<td>0.040</td>
<td>Garland</td>
<td>0.089</td>
<td>0.049</td>
<td>0.036</td>
<td>Newton</td>
<td>0.152</td>
<td>0.054</td>
<td>0.038</td>
</tr>
<tr>
<td>Ashley</td>
<td>0.108</td>
<td>0.058</td>
<td>0.058</td>
<td>Grant</td>
<td>0.108</td>
<td>0.044</td>
<td>0.031</td>
<td>Ouachita</td>
<td>0.108</td>
<td>0.068</td>
<td>0.037</td>
</tr>
<tr>
<td>Baxter</td>
<td>0.108</td>
<td>0.062</td>
<td>0.048</td>
<td>Greene</td>
<td>0.088</td>
<td>0.048</td>
<td>0.037</td>
<td>Perry</td>
<td>0.156</td>
<td>0.046</td>
<td>0.033</td>
</tr>
<tr>
<td>Benton</td>
<td>0.088</td>
<td>0.062</td>
<td>0.048</td>
<td>Hempstead</td>
<td>0.108</td>
<td>0.058</td>
<td>0.033</td>
<td>Phillips</td>
<td>0.108</td>
<td>0.076</td>
<td>0.054</td>
</tr>
<tr>
<td>Boone</td>
<td>0.109</td>
<td>0.062</td>
<td>0.046</td>
<td>Hot Spring</td>
<td>0.109</td>
<td>0.043</td>
<td>0.037</td>
<td>Pike</td>
<td>0.108</td>
<td>0.051</td>
<td>0.034</td>
</tr>
<tr>
<td>Bradley</td>
<td>0.108</td>
<td>0.054</td>
<td>0.039</td>
<td>Howard</td>
<td>0.108</td>
<td>0.044</td>
<td>0.043</td>
<td>Poinsett</td>
<td>0.108</td>
<td>0.058</td>
<td>0.032</td>
</tr>
<tr>
<td>Calhoun</td>
<td>0.153</td>
<td>0.048</td>
<td>0.044</td>
<td>Independence</td>
<td>0.088</td>
<td>0.046</td>
<td>0.035</td>
<td>Polk</td>
<td>0.108</td>
<td>0.068</td>
<td>0.049</td>
</tr>
<tr>
<td>Carroll</td>
<td>0.108</td>
<td>0.063</td>
<td>0.042</td>
<td>Izard</td>
<td>0.108</td>
<td>0.046</td>
<td>0.037</td>
<td>Pope</td>
<td>0.089</td>
<td>0.051</td>
<td>0.035</td>
</tr>
<tr>
<td>Chicot</td>
<td>0.154</td>
<td>0.088</td>
<td>0.058</td>
<td>Jackson</td>
<td>0.108</td>
<td>0.062</td>
<td>0.033</td>
<td>Prairie</td>
<td>0.154</td>
<td>0.062</td>
<td>0.036</td>
</tr>
<tr>
<td>Clark</td>
<td>0.109</td>
<td>0.054</td>
<td>0.036</td>
<td>Jefferson</td>
<td>0.088</td>
<td>0.058</td>
<td>0.037</td>
<td>Pulaski</td>
<td>0.088</td>
<td>0.044</td>
<td>0.031</td>
</tr>
<tr>
<td>Clay</td>
<td>0.108</td>
<td>0.069</td>
<td>0.039</td>
<td>Johnson</td>
<td>0.108</td>
<td>0.048</td>
<td>0.037</td>
<td>Randolph</td>
<td>0.108</td>
<td>0.051</td>
<td>0.040</td>
</tr>
<tr>
<td>Cleburne</td>
<td>0.108</td>
<td>0.046</td>
<td>0.033</td>
<td>Lafayette</td>
<td>0.153</td>
<td>0.062</td>
<td>0.051</td>
<td>St. Francis</td>
<td>0.108</td>
<td>0.051</td>
<td>0.041</td>
</tr>
<tr>
<td>Cleveland</td>
<td>0.153</td>
<td>0.046</td>
<td>0.039</td>
<td>Lawrence</td>
<td>0.108</td>
<td>0.051</td>
<td>0.031</td>
<td>Saline</td>
<td>0.088</td>
<td>0.041</td>
<td>0.033</td>
</tr>
<tr>
<td>Columbia</td>
<td>0.108</td>
<td>0.088</td>
<td>0.041</td>
<td>Lee</td>
<td>0.152</td>
<td>0.062</td>
<td>0.042</td>
<td>Scott</td>
<td>0.154</td>
<td>0.062</td>
<td>0.043</td>
</tr>
<tr>
<td>Conway</td>
<td>0.108</td>
<td>0.046</td>
<td>0.033</td>
<td>Lincoln</td>
<td>0.153</td>
<td>0.046</td>
<td>0.039</td>
<td>Searcy</td>
<td>0.154</td>
<td>0.062</td>
<td>0.035</td>
</tr>
<tr>
<td>Craighead</td>
<td>0.088</td>
<td>0.048</td>
<td>0.033</td>
<td>Little River</td>
<td>0.109</td>
<td>0.063</td>
<td>0.043</td>
<td>Sebastian</td>
<td>0.088</td>
<td>0.051</td>
<td>0.046</td>
</tr>
<tr>
<td>Crawford</td>
<td>0.089</td>
<td>0.046</td>
<td>0.037</td>
<td>Logan</td>
<td>0.109</td>
<td>0.054</td>
<td>0.036</td>
<td>Sevier</td>
<td>0.109</td>
<td>0.063</td>
<td>0.044</td>
</tr>
<tr>
<td>Crittenden</td>
<td>0.089</td>
<td>0.058</td>
<td>0.042</td>
<td>Lonoke</td>
<td>0.088</td>
<td>0.058</td>
<td>0.031</td>
<td>Sharp</td>
<td>0.108</td>
<td>0.044</td>
<td>0.040</td>
</tr>
<tr>
<td>Cross</td>
<td>0.108</td>
<td>0.048</td>
<td>0.038</td>
<td>Madison</td>
<td>0.109</td>
<td>0.054</td>
<td>0.034</td>
<td>Stone</td>
<td>0.154</td>
<td>0.054</td>
<td>0.037</td>
</tr>
<tr>
<td>Dallas</td>
<td>0.154</td>
<td>0.046</td>
<td>0.037</td>
<td>Marion</td>
<td>0.153</td>
<td>0.062</td>
<td>0.048</td>
<td>Union</td>
<td>0.089</td>
<td>0.077</td>
<td>0.048</td>
</tr>
<tr>
<td>Desha</td>
<td>0.108</td>
<td>0.088</td>
<td>0.054</td>
<td>Miller</td>
<td>0.089</td>
<td>0.077</td>
<td>0.046</td>
<td>Van Buren</td>
<td>0.108</td>
<td>0.046</td>
<td>0.040</td>
</tr>
<tr>
<td>Drew</td>
<td>0.108</td>
<td>0.058</td>
<td>0.044</td>
<td>Mississippi</td>
<td>0.088</td>
<td>0.088</td>
<td>0.038</td>
<td>Washington</td>
<td>0.089</td>
<td>0.046</td>
<td>0.040</td>
</tr>
<tr>
<td>Faulkner</td>
<td>0.089</td>
<td>0.054</td>
<td>0.031</td>
<td>Monroe</td>
<td>0.153</td>
<td>0.058</td>
<td>0.042</td>
<td>White</td>
<td>0.088</td>
<td>0.062</td>
<td>0.033</td>
</tr>
<tr>
<td>Franklin</td>
<td>0.108</td>
<td>0.044</td>
<td>0.034</td>
<td>Montgomery</td>
<td>0.154</td>
<td>0.063</td>
<td>0.037</td>
<td>Woodruff</td>
<td>0.153</td>
<td>0.046</td>
<td>0.036</td>
</tr>
<tr>
<td>Fulton</td>
<td>0.154</td>
<td>0.058</td>
<td>0.040</td>
<td>Nevada</td>
<td>0.153</td>
<td>0.058</td>
<td>0.038</td>
<td>Yell</td>
<td>0.108</td>
<td>0.069</td>
<td>0.035</td>
</tr>
</tbody>
</table>
As expected, sampling variability contributes to some initial differences in difficulty parameter estimates. However, as the bandwidth increases, local calibration samples increase in size and a smoothing effect is observed in these difficulty estimates. For instance, when calibration samples are based upon the 25 km bandwidth, the estimated difficulty parameters across the state for question 1 on the instrument have an observed range of 0.689 and interquartile range of 0.148. When the 75 km bandwidth is utilized, the same estimated difficulty parameters across the 75 local fittings have an observed range of 0.131 and interquartile range of 0.034. As Fotheringham et al. (2002, p. 211) noted, the bandwidth acts as a smoothing parameter with larger bandwidths corresponding to greater smoothing of the parameter estimate surface. The smoothing effect of the bandwidth can also be observed in the associated standard errors of the difficulty estimates. Local standard errors are highest at the 25 km bandwidth, especially for those counties allocated to have smaller local sample sizes of 250 examinees (see Table 5 and 8). The standard errors in the local model calibrations gradually decrease as the bandwidth increases. This smoothing property of the bandwidth was both anticipated for the proposed local IRT model and also demonstrated by these case study results of a moving window approach to 1PL IRT modeling. However, it should be noted that increases in bandwidth that correspond to smaller standard errors in estimation can also contribute to a reduction in power when identifying geographic-specific differences.

**2PL local IRT model results.** The localized approach to IRT modeling and DIF detection, utilizing a box-car kernel weighting scheme with three fixed bandwidths, was extended to the 2PL setting and applied to the Arkansas county 25-item instrument data, simulated to exhibit no regional DIF. Table 9 provides the global 2PL IRT discrimination estimates and associated standard errors for all 25 items as well as providing the 2PL IRT
discrimination estimates and associated standard errors for all 25 items averaged across the 75 local calibrations of the 2PL model for each of the three different fixed bandwidths. As the primary purpose of utilizing the 2PL local IRT model is to detect non-uniform regional DIF, which can be seen as a result of differing discrimination parameters between two compared groups (reference and focal), the estimated difficulty parameters are excluded from Table 9. More detailed results regarding the estimated discrimination parameters and associated errors of the 2PL local IRT model calibrations for each of the three different bandwidths are provided in Figure 25 and Tables 10 – 11.

Table 9. 2PL local IRT model summary – CS 1 estimated discrimination coefficients averaged across counties.

<table>
<thead>
<tr>
<th>Item (Question)</th>
<th>Global Model Estimated Discrim.</th>
<th>Global Model Standard Error</th>
<th>25 km Mean Estimated Discrim.</th>
<th>25 km Mean Standard Error</th>
<th>50 km Mean Estimated Discrim.</th>
<th>50 km Mean Standard Error</th>
<th>75 km Mean Estimated Discrim.</th>
<th>75 km Mean Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.989</td>
<td>0.016</td>
<td>1.001</td>
<td>0.149</td>
<td>0.989</td>
<td>0.073</td>
<td>0.990</td>
<td>0.051</td>
</tr>
<tr>
<td>2</td>
<td>0.330</td>
<td>0.012</td>
<td>0.340</td>
<td>0.112</td>
<td>0.337</td>
<td>0.055</td>
<td>0.330</td>
<td>0.038</td>
</tr>
<tr>
<td>3</td>
<td>1.090</td>
<td>0.017</td>
<td>1.101</td>
<td>0.159</td>
<td>1.097</td>
<td>0.077</td>
<td>1.088</td>
<td>0.054</td>
</tr>
<tr>
<td>4</td>
<td>1.624</td>
<td>0.023</td>
<td>1.637</td>
<td>0.215</td>
<td>1.614</td>
<td>0.104</td>
<td>1.615</td>
<td>0.073</td>
</tr>
<tr>
<td>5</td>
<td>1.701</td>
<td>0.030</td>
<td>1.720</td>
<td>0.283</td>
<td>1.709</td>
<td>0.136</td>
<td>1.703</td>
<td>0.095</td>
</tr>
<tr>
<td>6</td>
<td>0.485</td>
<td>0.014</td>
<td>0.487</td>
<td>0.128</td>
<td>0.486</td>
<td>0.062</td>
<td>0.489</td>
<td>0.044</td>
</tr>
<tr>
<td>7</td>
<td>0.648</td>
<td>0.014</td>
<td>0.652</td>
<td>0.131</td>
<td>0.649</td>
<td>0.064</td>
<td>0.644</td>
<td>0.045</td>
</tr>
<tr>
<td>8</td>
<td>0.271</td>
<td>0.012</td>
<td>0.269</td>
<td>0.110</td>
<td>0.267</td>
<td>0.054</td>
<td>0.270</td>
<td>0.038</td>
</tr>
<tr>
<td>9</td>
<td>1.613</td>
<td>0.028</td>
<td>1.648</td>
<td>0.263</td>
<td>1.637</td>
<td>0.128</td>
<td>1.614</td>
<td>0.088</td>
</tr>
<tr>
<td>10</td>
<td>1.251</td>
<td>0.018</td>
<td>1.255</td>
<td>0.169</td>
<td>1.258</td>
<td>0.083</td>
<td>1.252</td>
<td>0.058</td>
</tr>
<tr>
<td>11</td>
<td>0.993</td>
<td>0.017</td>
<td>1.002</td>
<td>0.154</td>
<td>0.996</td>
<td>0.075</td>
<td>0.995</td>
<td>0.053</td>
</tr>
<tr>
<td>12</td>
<td>1.359</td>
<td>0.021</td>
<td>1.367</td>
<td>0.197</td>
<td>1.360</td>
<td>0.096</td>
<td>1.362</td>
<td>0.067</td>
</tr>
<tr>
<td>13</td>
<td>1.602</td>
<td>0.027</td>
<td>1.650</td>
<td>0.261</td>
<td>1.614</td>
<td>0.123</td>
<td>1.598</td>
<td>0.085</td>
</tr>
<tr>
<td>14</td>
<td>0.901</td>
<td>0.015</td>
<td>0.915</td>
<td>0.142</td>
<td>0.902</td>
<td>0.069</td>
<td>0.903</td>
<td>0.048</td>
</tr>
<tr>
<td>15</td>
<td>1.657</td>
<td>0.024</td>
<td>1.677</td>
<td>0.225</td>
<td>1.659</td>
<td>0.109</td>
<td>1.657</td>
<td>0.076</td>
</tr>
<tr>
<td>16</td>
<td>1.477</td>
<td>0.023</td>
<td>1.489</td>
<td>0.210</td>
<td>1.480</td>
<td>0.102</td>
<td>1.477</td>
<td>0.071</td>
</tr>
<tr>
<td>17</td>
<td>1.387</td>
<td>0.022</td>
<td>1.425</td>
<td>0.214</td>
<td>1.406</td>
<td>0.102</td>
<td>1.394</td>
<td>0.071</td>
</tr>
<tr>
<td>18</td>
<td>1.103</td>
<td>0.017</td>
<td>1.106</td>
<td>0.154</td>
<td>1.103</td>
<td>0.076</td>
<td>1.104</td>
<td>0.053</td>
</tr>
<tr>
<td>19</td>
<td>0.888</td>
<td>0.017</td>
<td>0.889</td>
<td>0.152</td>
<td>0.877</td>
<td>0.074</td>
<td>0.884</td>
<td>0.052</td>
</tr>
<tr>
<td>20</td>
<td>1.636</td>
<td>0.047</td>
<td>1.726</td>
<td>0.493</td>
<td>1.659</td>
<td>0.219</td>
<td>1.648</td>
<td>0.151</td>
</tr>
<tr>
<td>21</td>
<td>1.699</td>
<td>0.024</td>
<td>1.731</td>
<td>0.224</td>
<td>1.703</td>
<td>0.107</td>
<td>1.700</td>
<td>0.075</td>
</tr>
<tr>
<td>22</td>
<td>1.467</td>
<td>0.021</td>
<td>1.485</td>
<td>0.194</td>
<td>1.475</td>
<td>0.094</td>
<td>1.466</td>
<td>0.066</td>
</tr>
<tr>
<td>23</td>
<td>1.581</td>
<td>0.022</td>
<td>1.600</td>
<td>0.203</td>
<td>1.587</td>
<td>0.099</td>
<td>1.582</td>
<td>0.069</td>
</tr>
<tr>
<td>24</td>
<td>0.490</td>
<td>0.013</td>
<td>0.491</td>
<td>0.115</td>
<td>0.490</td>
<td>0.056</td>
<td>0.491</td>
<td>0.039</td>
</tr>
<tr>
<td>25</td>
<td>0.313</td>
<td>0.012</td>
<td>0.317</td>
<td>0.111</td>
<td>0.314</td>
<td>0.054</td>
<td>0.313</td>
<td>0.038</td>
</tr>
</tbody>
</table>
Figure 25. 2PL local IRT model discrimination estimates and standard errors for CS 1.
Table 10. 2PL local IRT model summary – CS 1 estimated discrimination coefficients for question 1 by county.

<table>
<thead>
<tr>
<th>County</th>
<th>25 km Estimated Discrim.</th>
<th>50 km Estimated Discrim.</th>
<th>75 km Estimated Discrim.</th>
<th>County</th>
<th>25 km Estimated Discrim.</th>
<th>50 km Estimated Discrim.</th>
<th>75 km Estimated Discrim.</th>
<th>County</th>
<th>25 km Estimated Discrim.</th>
<th>50 km Estimated Discrim.</th>
<th>75 km Estimated Discrim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arkansas</td>
<td>0.985</td>
<td>0.978</td>
<td>0.973</td>
<td>Garland</td>
<td>0.953</td>
<td>0.998</td>
<td>1.014</td>
<td>Newton</td>
<td>1.041</td>
<td>1.039</td>
<td>1.041</td>
</tr>
<tr>
<td>Ashley</td>
<td>1.022</td>
<td>0.959</td>
<td>0.959</td>
<td>Grant</td>
<td>0.940</td>
<td>0.965</td>
<td>0.945</td>
<td>Ouachita</td>
<td>0.768</td>
<td>0.957</td>
<td>0.915</td>
</tr>
<tr>
<td>Baxter</td>
<td>0.864</td>
<td>0.938</td>
<td>1.010</td>
<td>Greene</td>
<td>1.050</td>
<td>0.967</td>
<td>0.971</td>
<td>Perry</td>
<td>1.179</td>
<td>1.018</td>
<td>0.988</td>
</tr>
<tr>
<td>Benton</td>
<td>0.972</td>
<td>1.001</td>
<td>1.059</td>
<td>Hempstead</td>
<td>1.155</td>
<td>0.934</td>
<td>1.006</td>
<td>Phillips</td>
<td>0.925</td>
<td>1.017</td>
<td>1.002</td>
</tr>
<tr>
<td>Boone</td>
<td>1.087</td>
<td>1.022</td>
<td>1.015</td>
<td>Hot Spring</td>
<td>0.902</td>
<td>0.966</td>
<td>1.011</td>
<td>Pike</td>
<td>0.886</td>
<td>1.018</td>
<td>0.988</td>
</tr>
<tr>
<td>Bradley</td>
<td>0.929</td>
<td>0.916</td>
<td>0.948</td>
<td>Howard</td>
<td>1.065</td>
<td>1.038</td>
<td>1.055</td>
<td>Poinsett</td>
<td>0.915</td>
<td>1.001</td>
<td>0.970</td>
</tr>
<tr>
<td>Calhoun</td>
<td>1.149</td>
<td>0.929</td>
<td>0.907</td>
<td>Independence</td>
<td>1.156</td>
<td>1.034</td>
<td>1.072</td>
<td>Polk</td>
<td>1.210</td>
<td>0.988</td>
<td>1.058</td>
</tr>
<tr>
<td>Carroll</td>
<td>1.183</td>
<td>1.136</td>
<td>1.061</td>
<td>Izard</td>
<td>1.196</td>
<td>1.048</td>
<td>1.010</td>
<td>Pope</td>
<td>0.976</td>
<td>1.055</td>
<td>1.025</td>
</tr>
<tr>
<td>Chicot</td>
<td>0.726</td>
<td>1.065</td>
<td>0.992</td>
<td>Jackson</td>
<td>1.021</td>
<td>1.128</td>
<td>1.054</td>
<td>Prairie</td>
<td>1.077</td>
<td>0.979</td>
<td>0.974</td>
</tr>
<tr>
<td>Clark</td>
<td>1.131</td>
<td>0.959</td>
<td>0.980</td>
<td>Jefferson</td>
<td>0.923</td>
<td>0.942</td>
<td>0.951</td>
<td>Pulaski</td>
<td>1.017</td>
<td>0.930</td>
<td>0.983</td>
</tr>
<tr>
<td>Clay</td>
<td>0.945</td>
<td>0.994</td>
<td>0.964</td>
<td>Johnson</td>
<td>0.951</td>
<td>0.984</td>
<td>1.010</td>
<td>Randolph</td>
<td>1.059</td>
<td>0.959</td>
<td>0.999</td>
</tr>
<tr>
<td>Cleburne</td>
<td>1.301</td>
<td>1.161</td>
<td>1.053</td>
<td>Lafayette</td>
<td>1.195</td>
<td>0.984</td>
<td>0.917</td>
<td>St. Francis</td>
<td>0.685</td>
<td>0.943</td>
<td>0.971</td>
</tr>
<tr>
<td>Cleveland</td>
<td>0.968</td>
<td>0.947</td>
<td>0.942</td>
<td>Lawrence</td>
<td>0.904</td>
<td>0.954</td>
<td>1.026</td>
<td>Saline</td>
<td>0.840</td>
<td>0.946</td>
<td>0.964</td>
</tr>
<tr>
<td>Columbia</td>
<td>0.841</td>
<td>1.019</td>
<td>0.922</td>
<td>Lee</td>
<td>1.110</td>
<td>1.031</td>
<td>0.956</td>
<td>Scott</td>
<td>0.965</td>
<td>0.941</td>
<td>0.985</td>
</tr>
<tr>
<td>Conway</td>
<td>1.183</td>
<td>1.042</td>
<td>1.000</td>
<td>Lincoln</td>
<td>0.960</td>
<td>0.905</td>
<td>0.909</td>
<td>Searcy</td>
<td>1.065</td>
<td>0.980</td>
<td>1.007</td>
</tr>
<tr>
<td>Craighead</td>
<td>0.911</td>
<td>0.973</td>
<td>1.002</td>
<td>Little River</td>
<td>0.918</td>
<td>1.127</td>
<td>1.044</td>
<td>Sebastian</td>
<td>1.028</td>
<td>0.943</td>
<td>0.946</td>
</tr>
<tr>
<td>Crawford</td>
<td>1.083</td>
<td>0.909</td>
<td>0.953</td>
<td>Logan</td>
<td>1.135</td>
<td>0.979</td>
<td>0.992</td>
<td>Sevier</td>
<td>1.046</td>
<td>1.127</td>
<td>1.038</td>
</tr>
<tr>
<td>Crittenden</td>
<td>1.051</td>
<td>0.921</td>
<td>0.947</td>
<td>Lonoke</td>
<td>0.945</td>
<td>0.929</td>
<td>0.963</td>
<td>Sharp</td>
<td>0.849</td>
<td>1.034</td>
<td>1.067</td>
</tr>
<tr>
<td>Cross</td>
<td>0.969</td>
<td>0.953</td>
<td>0.983</td>
<td>Madison</td>
<td>0.851</td>
<td>1.013</td>
<td>1.007</td>
<td>Stone</td>
<td>1.273</td>
<td>0.980</td>
<td>1.007</td>
</tr>
<tr>
<td>Dallas</td>
<td>0.743</td>
<td>0.972</td>
<td>0.932</td>
<td>Marion</td>
<td>1.215</td>
<td>0.976</td>
<td>0.969</td>
<td>Union</td>
<td>1.101</td>
<td>0.880</td>
<td>0.935</td>
</tr>
<tr>
<td>Desha</td>
<td>0.988</td>
<td>0.967</td>
<td>0.934</td>
<td>Miller</td>
<td>0.962</td>
<td>0.941</td>
<td>0.959</td>
<td>Van Buren</td>
<td>0.949</td>
<td>1.069</td>
<td>1.020</td>
</tr>
<tr>
<td>Drew</td>
<td>0.866</td>
<td>0.945</td>
<td>0.938</td>
<td>Mississippi</td>
<td>0.938</td>
<td>0.911</td>
<td>0.927</td>
<td>Washington</td>
<td>0.953</td>
<td>1.009</td>
<td>0.997</td>
</tr>
<tr>
<td>Faulkner</td>
<td>1.054</td>
<td>0.938</td>
<td>1.007</td>
<td>Monroe</td>
<td>1.037</td>
<td>0.973</td>
<td>0.985</td>
<td>White</td>
<td>0.906</td>
<td>1.034</td>
<td>1.009</td>
</tr>
<tr>
<td>Franklin</td>
<td>0.899</td>
<td>0.931</td>
<td>0.978</td>
<td>Montgomery</td>
<td>0.775</td>
<td>1.037</td>
<td>1.009</td>
<td>Woodruff</td>
<td>1.381</td>
<td>1.057</td>
<td>1.077</td>
</tr>
<tr>
<td>Fulton</td>
<td>0.820</td>
<td>0.954</td>
<td>1.031</td>
<td>Nevada</td>
<td>0.888</td>
<td>0.886</td>
<td>0.973</td>
<td>Yell</td>
<td>1.165</td>
<td>1.075</td>
<td>1.029</td>
</tr>
</tbody>
</table>
Table 11. 2PL local IRT model summary – CS 1 standard errors for estimated discrimination coefficients for question 1 by county.

<table>
<thead>
<tr>
<th>County</th>
<th>25 km Estimated Discrim.</th>
<th>50 km Estimated Discrim.</th>
<th>75 km Estimated Discrim.</th>
<th>County</th>
<th>25 km Estimated Discrim.</th>
<th>50 km Estimated Discrim.</th>
<th>75 km Estimated Discrim.</th>
<th>County</th>
<th>25 km Estimated Discrim.</th>
<th>50 km Estimated Discrim.</th>
<th>75 km Estimated Discrim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arkansas</td>
<td>0.136</td>
<td>0.098</td>
<td>0.049</td>
<td>Garland</td>
<td>0.108</td>
<td>0.059</td>
<td>0.047</td>
<td>Newton</td>
<td>0.206</td>
<td>0.061</td>
<td>0.049</td>
</tr>
<tr>
<td>Ashley</td>
<td>0.142</td>
<td>0.068</td>
<td>0.068</td>
<td>Grant</td>
<td>0.137</td>
<td>0.053</td>
<td>0.040</td>
<td>Ouachita</td>
<td>0.124</td>
<td>0.073</td>
<td>0.048</td>
</tr>
<tr>
<td>Baxter</td>
<td>0.130</td>
<td>0.072</td>
<td>0.060</td>
<td>Greene</td>
<td>0.116</td>
<td>0.061</td>
<td>0.047</td>
<td>Perry</td>
<td>0.205</td>
<td>0.057</td>
<td>0.042</td>
</tr>
<tr>
<td>Benton</td>
<td>0.113</td>
<td>0.099</td>
<td>0.071</td>
<td>Hempstead</td>
<td>0.150</td>
<td>0.068</td>
<td>0.045</td>
<td>Phillips</td>
<td>0.133</td>
<td>0.101</td>
<td>0.075</td>
</tr>
<tr>
<td>Boone</td>
<td>0.146</td>
<td>0.075</td>
<td>0.057</td>
<td>Hot Spring</td>
<td>0.132</td>
<td>0.056</td>
<td>0.047</td>
<td>Pike</td>
<td>0.139</td>
<td>0.067</td>
<td>0.044</td>
</tr>
<tr>
<td>Bradley</td>
<td>0.136</td>
<td>0.063</td>
<td>0.050</td>
<td>Howard</td>
<td>0.148</td>
<td>0.061</td>
<td>0.057</td>
<td>Poinsett</td>
<td>0.134</td>
<td>0.075</td>
<td>0.041</td>
</tr>
<tr>
<td>Calhoun</td>
<td>0.215</td>
<td>0.060</td>
<td>0.057</td>
<td>Independence</td>
<td>0.126</td>
<td>0.061</td>
<td>0.047</td>
<td>Polk</td>
<td>0.156</td>
<td>0.079</td>
<td>0.059</td>
</tr>
<tr>
<td>Carroll</td>
<td>0.154</td>
<td>0.086</td>
<td>0.059</td>
<td>Izard</td>
<td>0.153</td>
<td>0.064</td>
<td>0.048</td>
<td>Pope</td>
<td>0.112</td>
<td>0.067</td>
<td>0.045</td>
</tr>
<tr>
<td>Chicot</td>
<td>0.173</td>
<td>0.091</td>
<td>0.074</td>
<td>Jackson</td>
<td>0.140</td>
<td>0.080</td>
<td>0.045</td>
<td>Prairie</td>
<td>0.212</td>
<td>0.065</td>
<td>0.045</td>
</tr>
<tr>
<td>Clark</td>
<td>0.151</td>
<td>0.065</td>
<td>0.046</td>
<td>Jefferson</td>
<td>0.111</td>
<td>0.072</td>
<td>0.043</td>
<td>Pulaski</td>
<td>0.115</td>
<td>0.055</td>
<td>0.038</td>
</tr>
<tr>
<td>Clay</td>
<td>0.136</td>
<td>0.079</td>
<td>0.050</td>
<td>Johnson</td>
<td>0.137</td>
<td>0.062</td>
<td>0.047</td>
<td>Randolph</td>
<td>0.146</td>
<td>0.065</td>
<td>0.049</td>
</tr>
<tr>
<td>Cleburne</td>
<td>0.164</td>
<td>0.082</td>
<td>0.047</td>
<td>Lafayette</td>
<td>0.216</td>
<td>0.081</td>
<td>0.061</td>
<td>St. Francis</td>
<td>0.122</td>
<td>0.065</td>
<td>0.052</td>
</tr>
<tr>
<td>Cleveland</td>
<td>0.204</td>
<td>0.053</td>
<td>0.049</td>
<td>Lawrence</td>
<td>0.133</td>
<td>0.068</td>
<td>0.040</td>
<td>Saline</td>
<td>0.106</td>
<td>0.051</td>
<td>0.040</td>
</tr>
<tr>
<td>Columbia</td>
<td>0.130</td>
<td>0.101</td>
<td>0.053</td>
<td>Lee</td>
<td>0.214</td>
<td>0.091</td>
<td>0.050</td>
<td>Scott</td>
<td>0.194</td>
<td>0.078</td>
<td>0.054</td>
</tr>
<tr>
<td>Conway</td>
<td>0.155</td>
<td>0.061</td>
<td>0.042</td>
<td>Lincoln</td>
<td>0.192</td>
<td>0.063</td>
<td>0.052</td>
<td>Searcy</td>
<td>0.204</td>
<td>0.074</td>
<td>0.046</td>
</tr>
<tr>
<td>Craighead</td>
<td>0.110</td>
<td>0.065</td>
<td>0.041</td>
<td>Little River</td>
<td>0.135</td>
<td>0.085</td>
<td>0.059</td>
<td>Sebastian</td>
<td>0.114</td>
<td>0.061</td>
<td>0.058</td>
</tr>
<tr>
<td>Crawford</td>
<td>0.119</td>
<td>0.060</td>
<td>0.047</td>
<td>Logan</td>
<td>0.148</td>
<td>0.069</td>
<td>0.046</td>
<td>Sevier</td>
<td>0.141</td>
<td>0.085</td>
<td>0.061</td>
</tr>
<tr>
<td>Crittenden</td>
<td>0.118</td>
<td>0.079</td>
<td>0.056</td>
<td>Lonoke</td>
<td>0.111</td>
<td>0.064</td>
<td>0.039</td>
<td>Sharp</td>
<td>0.131</td>
<td>0.058</td>
<td>0.051</td>
</tr>
<tr>
<td>Cross</td>
<td>0.138</td>
<td>0.061</td>
<td>0.049</td>
<td>Madison</td>
<td>0.131</td>
<td>0.066</td>
<td>0.044</td>
<td>Stone</td>
<td>0.221</td>
<td>0.075</td>
<td>0.050</td>
</tr>
<tr>
<td>Dallas</td>
<td>0.172</td>
<td>0.059</td>
<td>0.045</td>
<td>Marion</td>
<td>0.224</td>
<td>0.080</td>
<td>0.059</td>
<td>Union</td>
<td>0.119</td>
<td>0.094</td>
<td>0.068</td>
</tr>
<tr>
<td>Desha</td>
<td>0.142</td>
<td>0.139</td>
<td>0.072</td>
<td>Miller</td>
<td>0.114</td>
<td>0.087</td>
<td>0.056</td>
<td>Van Buren</td>
<td>0.138</td>
<td>0.065</td>
<td>0.052</td>
</tr>
<tr>
<td>Drew</td>
<td>0.131</td>
<td>0.073</td>
<td>0.053</td>
<td>Mississippi</td>
<td>0.110</td>
<td>0.110</td>
<td>0.049</td>
<td>Washington</td>
<td>0.112</td>
<td>0.063</td>
<td>0.055</td>
</tr>
<tr>
<td>Faulkner</td>
<td>0.114</td>
<td>0.064</td>
<td>0.041</td>
<td>Monroe</td>
<td>0.202</td>
<td>0.065</td>
<td>0.051</td>
<td>White</td>
<td>0.110</td>
<td>0.090</td>
<td>0.043</td>
</tr>
<tr>
<td>Franklin</td>
<td>0.134</td>
<td>0.058</td>
<td>0.042</td>
<td>Montgomery</td>
<td>0.180</td>
<td>0.083</td>
<td>0.048</td>
<td>Woodruff</td>
<td>0.239</td>
<td>0.057</td>
<td>0.047</td>
</tr>
<tr>
<td>Fulton</td>
<td>0.184</td>
<td>0.073</td>
<td>0.054</td>
<td>Nevada</td>
<td>0.188</td>
<td>0.077</td>
<td>0.049</td>
<td>Yell</td>
<td>0.152</td>
<td>0.092</td>
<td>0.042</td>
</tr>
</tbody>
</table>
In the 2PL local IRT modeling of the Arkansas county data, simulated to have only one latent spatial subgroup and to display no regional DIF, the bandwidth functions once more as a smoothing parameter. As the selected bandwidth increases, variability in the observed discrimination estimates decreases. For instance, calibrating the local models with a 25 km bandwidth results in estimated discrimination parameters across the state on question 1 to have an observed range of 0.696 and interquartile range of 0.172. When the 75 km bandwidth is utilized for local model calibrations, the same estimated discrimination parameters across the 75 local fittings have an observed range of 0.170 and interquartile range of 0.054. The smoothing effect of the bandwidth, previously noted in the 1PL local IRT setting, can also be observed in this 2PL local IRT setting in both the smoothing of estimated model parameters and their associated standard errors. It is of particular note that the local standard errors are highest at the 25 km bandwidth in accordance with local sample sizes (see Table 5 and 11). Furthermore, it is noteworthy that, though the standard errors in the local model calibrations gradually decrease as the bandwidth increases, standard errors for areal units located on the border remain higher relative to those for areal units in the center of the study area even as the bandwidth increases. This is an artifact of the neighborhood structure and demonstrates the spatial connectedness of centrally located areal units in contrast to the spatial disconnectedness of bordering areal units. This smoothing property of the bandwidth and the increased local standard errors observed for border calibration sites does not break with the expectation from the literature.

Regional DIF detection. Traditional DIF detection techniques require that only two groups be considered, the focal group and the reference group. Consequently, typical approaches to DIF detection in a multiple group setting across a spatial area of interest rely on either the application of several pairwise comparisons or naïve comparisons based upon cardinal direction
(Svetina & Rutkowski, 2014). While it is recognized that comparisons based only upon cardinal direction might fail to detect regional DIF owing to the arbitrary aggregation of areal units, the application of pairwise comparisons for multiple group DIF detection can be overwhelming. For instance, in the current case study data were provided for 75 counties resulting in 2775 possible pairwise comparisons. To reduce the number of pairwise comparisons conducted in DIF analyses, previous approaches that artificially created two groups were proposed (Hambleton & Kanjee, 1995). One of these approaches defined focal and reference groups so as to compare one group against the aggregate of the other groups, which would result in 75 pairwise comparisons in the context of this simulated case study. Some of these more traditional approaches to regional DIF detection will be compared here to the proposed method.

In order to detect regional DIF utilizing the localized approach to IRT modeling, it is suggested in this current work that estimated item difficulty parameters (in the 1PL setting) or discrimination parameters (in the 2PL setting) be treated as fixed values and assessed for spatial nonstationarity and local effects such as spatial clustering. Utilizing Local Moran’s I and associated one-tailed significance testing, regions identified as having significant spatial clustering can serve as spatial subgroups for use in subsequent DIF analyses. The results of significance testing for spatial clustering in this case study are provided in Figure 26 and Figure 27. Figure 26 provides the p-values and associated Local Moran’s I values for each of the three different bandwidths using the estimated difficulty parameters obtained from the 1PL local IRT model calibrations, which would be used for subgroup construction to assess uniform regional DIF. Figure 27, similarly, provides the p-values and associated Local Moran’s I values for each of the three different bandwidths using the estimated discrimination parameters obtained from
Figure 26. 1PL local IRT model local Moran’s I results for CS 1.
Figure 27. 2PL local IRT model local Moran’s I results for CS 1.
the 2PL local IRT model calibrations, which would be used for subgroup construction to assess non-uniform regional DIF. Note that owing to the fact that the sign of Local Moran’s I is more interpretable than the magnitude of the value itself since it is not bounded on [-1,+1] as previously discussed, only mappings of p-values for one-tailed significance testing of spatial clustering will be provided in subsequent case studies.

Traditional DIF detection methods using a non-proximal approach such as the Mantel-Haenszel chi-square, logistic regression, and Raju’s Exact Unsigned Area (EUA) methods reveal no DIF (uniform or non-uniform) detected on question 1 of the instrument between North and South or between East and West subgroups for the sets of simulated data at a significance level of 0.05. This would, of course, not be incorrect since this case study was specifically designed so that there was only one latent spatial subgroup and no regional DIF present. If, however, 75 pairwise comparisons are made between each county and the aggregate of the remaining counties utilizing three traditional DIF detection methods (MH, Logistic Regression, and Raju’s EUA), several counties are reported as having question 1 flagged for significant DIF (see Table 12). The results indicate a fairly well controlled type I error rate for the 1PL condition (.040 to .053) with inflated type I error for the 2PL condition (.067 to .160).

<table>
<thead>
<tr>
<th>Model Type</th>
<th>MH</th>
<th>Logistic Regression</th>
<th>Raju’s Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1PL</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>2PL</td>
<td>7</td>
<td>5</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 12. Number of counties flagged for DIF on question 1 using pairwise comparisons.

While this number of pairwise comparisons will not be conducted in subsequent case studies, it is provided for illustrative purpose in the current case study, as it is representative of the
potential for type I error when utilizing multiple pairwise comparisons of traditional item-level DIF techniques with spatial subgroups defined by political boundaries. Further, it can be compared to the proposed method more directly.

Setting again a significance level of 0.05, the p-values ascertained from one-tailed Local Moran’s I significance tests for spatial clustering of the 1PL local IRT model estimated difficulty parameters are mapped more clearly in Figure 28. Counties in red suggest evidence of significant spatial clustering at each of the fixed bandwidths. The fourth map, however, displays counties in red that have demonstrated significant spatial clustering at all of the three fixed bandwidths.

*Figure 28. 1PL local IRT model spatial subgroups utilizing local Moran’s I for CS 1.*
selected. These four counties (Scott, Polk, Cleaveland, and Bradley) might comprise one or two latent spatial subgroups. Further investigation of the estimated difficulty parameters for these counties (see Table 7) indicate that the four counties comprise two spatial subgroups, each of which can be assessed for regional uniform DIF with a traditional pairwise comparison DIF technique (Subgroup 1 vs. No DIF Counties, Subgroup 2 vs. No DIF Counties, and Subgroup 1 vs. Subgroup 2) or could each be flagged for potential regional uniform DIF as is, which would yield comparable results in the 1PL setting to the number of counties flagged for question 1 DIF using 75 pairwise comparisons (see Table 12).

Applying the same procedure to the 2PL local IRT model estimated discrimination parameters, mapped results of one-tailed Local Moran’s I significance tests for spatial clustering are presented in Figure 29. Counties in red suggest evidence of significant spatial clustering at each of the fixed bandwidths. Focusing once more on the fourth map, there is only one county (Cleburne) that has demonstrated significant spatial clustering at all of the three fixed bandwidths selected. Given the relative sensitivity of the spatial clustering at different user-specified bandwidths, this result would suggest that there does not appear to be regional non-uniform DIF present for question 1 in the simulated sample data. This result is very promising, especially when comparisons are made to the many pairwise comparisons that might be made for regional DIF detection (see Table 12).

The primary aim of this first case study was to investigate the impact of bandwidth selection, local (i.e., calibration) sample size, and model type on the proposed method results when only one latent spatial subgroup is present and, as such, no regional DIF should be detected. Results were presented for both the 1PL and the 2PL local IRT models. This first case study demonstrated the smoothing effect of the bandwidth, revealed the impact that calibration
sample sizes have on both parameter estimates and local standard errors, and suggested that the localized approach to IRT modeling is a potentially comparable or superior approach to both uniform and non-uniform regional DIF detection. This case study also introduced the reader to the methodology itself and concepts such as spatial stationarity in context.

Figure 29. 2PL local IRT model spatial subgroups utilizing local Moran’s I for CS 1.
Case Study 2 – Simulated Data Exhibiting Regional DIF with Spatial Clustering

This second case study applied the local IRT modeling approach for regional DIF detection to data that has been simulated for each of the 75 counties in the state of Arkansas. Data were simulated as in the previous case study however certain factors were manipulated such as DIF type (uniform and non-uniform) and DIF magnitude, which was modified by altering the area between IRFs of reference and focal group member counties to be approximately 0.40, 0.60, and 0.80 as previously discussed. The number of examinees in each county was once more allocated in accordance with the relative size of the 2016-2017 Arkansas school enrollments (see Figure 18 and Table 5) and the centroids of each county serve as the regression points for local model calibration in both the 1PL and 2PL settings (see Figure 17). As before, Local IRT models were calibrated at each of three fixed bandwidths of 25 km, 50 km, and 75 km (see Figure 23).

It must be noted that 75 local calibrations of IRT fitted models (both 1PL and 2PL) for an instrument of 25 items at three fixed bandwidths produced 33,750 distinct coefficient and standard error local estimates (excluding the unit discrimination parameter estimate constrained in the 1PL fitted models). This number would exceed 100,000 local estimates for the current case study as data were simulated at three different levels of DIF magnitude. Consequently, only relevant and exemplar results are provided below. Focus will be paid to the results for the first item on the instrument, as this item was chosen during the simulation process to exhibit DIF.

The primary aim of this second case study was to demonstrate the benefit of the proposed method when a spatial pattern exists, specifically when a spiral spatial clustering pattern centered at the two major MSAs in Arkansas was present. The hypothesized benefits of the method were investigated at different levels of DIF type and DIF magnitude, with comparisons made to traditional, non-proximal DIF detection methods. The results will be presented for the 1PL local
IRT model and the 2PL local IRT model across all manipulated settings before discussing the primary findings and possible generalizations derived from this case study.

**1PL local IRT model results.** The localized approach to IRT modeling and DIF detection, utilizing a box-car kernel weighting scheme with three fixed bandwidths, was applied to the Arkansas county 25-item instrument data, simulated to exhibit a small, moderate, or large amount of regional DIF based upon inclusion or lack thereof in one of two MSAs (Fayetteville-Springdale-Rogers MSA or Little Rock-North Little Rock-Conway MSA). The counties belonging to one of these MSAs were previously displayed in Figure 19. Tables 13 - 14 provide basic summary statistics for question 1 difficulty estimates and associated standard errors of the 75 local calibrations of the 1PL IRT model for each of the three fixed bandwidths and three DIF magnitudes. Note that, for all 1PL local model calibrations, the discrimination parameter was constrained to one. Furthermore, Tables 13 - 14 also provide 2PL local IRT model results.

The local calibrations of the 1PL IRT models for each of the three bandwidths and three DIF magnitudes are further displayed by providing surface mappings of county-specific difficulty estimates for question 1 of the instrument (see Figure 30).

As in the previous case study, several features of local spatial modeling can be observed. For instance, the bandwidth functions once more as a smoothing parameter, with difficulty estimate surfaces becoming more smooth as the bandwidth increases no matter the level of DIF magnitude. This smoothing feature of the bandwidth is visually apparent in Figure 30 and can also be discerned from Tables 13 - 14 by observing the decreasing range of the local calibration difficulty estimates at each DIF magnitude level across the three fixed bandwidths. In addition, while neither presented for visualization in Figure 30 nor apparent from the tables, local standard errors are again highest at the 25 km bandwidth in accordance with the local sample sizes (see
Table 13. *Local IRT model summary – CS 2 estimated difficulty and discrimination coefficients for question 1 across all 75 Arkansas counties.*

<table>
<thead>
<tr>
<th>Model</th>
<th>Coefficient Estimate Summarized</th>
<th>Simulated Area Between IRFs</th>
<th>25 km Bandwidth</th>
<th>50 km Bandwidth</th>
<th>75 km Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Min.</td>
<td>Max.</td>
<td>Mean SE</td>
</tr>
<tr>
<td>1PL</td>
<td>Difficulty</td>
<td>0.40</td>
<td>0.025</td>
<td>-0.338</td>
<td>0.628</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>0.039</td>
<td>-0.338</td>
<td>0.686</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>0.062</td>
<td>-0.338</td>
<td>0.995</td>
</tr>
<tr>
<td>2PL</td>
<td>Discrimination</td>
<td>0.40</td>
<td>0.982</td>
<td>0.612</td>
<td>1.381</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>0.967</td>
<td>0.567</td>
<td>1.381</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>0.964</td>
<td>0.559</td>
<td>1.381</td>
</tr>
</tbody>
</table>
Table 14. Local IRT model summary – CS 2 estimated difficulty and discrimination coefficients for question 1 across reference and focal Arkansas counties.

<table>
<thead>
<tr>
<th>Model</th>
<th>Coefficient Estimate Summarized</th>
<th>Simulated Area Between IRFs</th>
<th>25 km Bandwidth</th>
<th>50 km Bandwidth</th>
<th>75 km Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Min.</td>
<td>Max.</td>
<td>Mean</td>
</tr>
<tr>
<td>1PL</td>
<td>Reference Group Difficulty</td>
<td>0.40</td>
<td>-0.032</td>
<td>-0.338</td>
<td>0.351</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>-0.032</td>
<td>-0.338</td>
<td>0.351</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>-0.032</td>
<td>-0.338</td>
<td>0.351</td>
</tr>
<tr>
<td></td>
<td>Focal Group Difficulty</td>
<td>0.40</td>
<td>0.439</td>
<td>0.313</td>
<td>0.628</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>0.554</td>
<td>0.367</td>
<td>0.686</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>0.745</td>
<td>0.444</td>
<td>0.995</td>
</tr>
<tr>
<td>2PL</td>
<td>Reference Group Difficulty</td>
<td>0.40</td>
<td>1.005</td>
<td>0.685</td>
<td>1.381</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>1.005</td>
<td>0.685</td>
<td>1.381</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>1.005</td>
<td>0.685</td>
<td>1.381</td>
</tr>
<tr>
<td></td>
<td>Focal Group Difficulty</td>
<td>0.40</td>
<td>0.814</td>
<td>0.612</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>0.689</td>
<td>0.567</td>
<td>0.849</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>0.661</td>
<td>0.559</td>
<td>0.820</td>
</tr>
</tbody>
</table>
Figure 30. CS 2 IPI local IRT model difficulty estimates across fixed bandwidths and three DIF magnitudes (small, moderate, large).
Table 5) and gradually decrease as the bandwidth increases. Furthermore, error in local estimations is more apparent for areal units located at the edges of the study region, where specified neighborhood structures generate smaller calibration sample sizes.

The Arkansas county data for the current simulated case study was created to display two latent spatial subgroups and observable regional DIF. The two latent spatial subgroups, one subgroup representing counties that are a part of the MSAs and one subgroup representing the remaining Arkansas counties. Most striking about Figure 30 is the fact that the spiral spatial clustering pattern centered around the two MSAs in northwest and central Arkansas are very discernable, especially as the magnitude of DIF increases. The smallest simulated DIF magnitude is more difficult to detect visually, especially as the bandwidth increases and difficulty estimates experience the smoothing effect of that increase. However, at moderate or large DIF magnitudes, spatial nonstationarity in local difficulty parameter estimates is visually noticeable even as the bandwidth increases to 75 km. Based upon these 1PL local IRT modeling results, there is value in visualizing difficulty estimates across space, especially in the presence of spatial clustering patterns and moderate to large DIF magnitudes.

2PL local IRT model results. The 2PL model of the proposed method for regional DIF detection was applied once more to Arkansas county 25-item instrument data, simulated to exhibit a small, moderate, or large amount of regional non-uniform DIF. Tables 13-14, previously presented, provide basic summary statistics for question 1 discrimination estimates and associated standard errors of the 75 local calibrations of the 2PL IRT model for each of the three fixed bandwidths and three DIF magnitudes. Difficulty parameters were excluded from Tables 13 - 14 for the 2PL setting. Surface mappings of county-specific discrimination estimates for question 1 of the instrument are provided in Figure 31.
Figure 31. CS 2 2PL local IRT model discrimination estimates across fixed bandwidths and DIF magnitudes (small, moderate, large).
As in both the previous case study and the 1PL local IRT modeling of the data exhibiting uniform regional DIF in the current case study, the bandwidth again serves as a smoothing parameter and local standard errors are still related to calibration sample size and spatial connectivity no matter the level of DIF magnitude or the type of DIF induced. This smoothing feature of the bandwidth is evident in Figure 31, though less so than in the uniform DIF setting so discernable when utilizing a 1PL local IRT modeling approach. The smoothing effect of the bandwidth can also be observed from Tables 13 - 14.

The 2PL local IRT approach was applied to simulated data in the current case study that displayed non-uniform regional DIF in two latent spatial subgroups based upon MSA membership. Similar to the 1PL local IRT approach and the simulated uniform regional DIF data, a spatial clustering pattern does emerge. However, this clustering pattern is less apparent visually than in the previous setting (see Figure 31). Even as the magnitude of DIF increases, spatial subgroups do not appear any more distinct. Additionally, in this non-uniform DIF setting, larger bandwidths seem to be necessary for visual identification of the central MSA region with the northwest Arkansas MSA region less apparent. This lack of visual inference could be due to either sampling variability or graphical decisions such as the fill scale utilized and the color palette applied. Further investigation of differences in estimated discrimination parameters and the ability of the 2PL local IRT model to detect non-uniform regional DIF will be discussed in the forthcoming section.

**Regional DIF detection.** The current case study was presented to explore the benefits of local IRT modeling for regional DIF detection, especially when a latent spatial structure is present that might render traditional DIF techniques comparing regions North to South or East to West inadequate. It was hypothesized that the spatial clustering pattern in the current case study
would make detection of regional DIF particularly challenging for non-proximal pairwise comparisons based upon cardinal direction. To investigate this, three traditional DIF detection methods (Mantel-Haenszel chi-square, logistic regression, and Raju’s EUA) were applied to the various simulated datasets, comparing subgroups North to South and East to West for each model type and DIF magnitude setting. With only one exception, these traditional methods were unable to detect DIF (uniform or non-uniform) for question 1 of the instrument at a significance level of 0.05. To summarize these non-proximal, pairwise comparisons Table 15 is provided, where ‘X’ indicates the first instrument item was not flagged for DIF and ‘✓’ indicates the first instrument item was flagged for DIF by the method. These findings are, of course, not correct since this case study was specifically designed so that there was regional DIF present. The common use of traditional DIF techniques applied to spatial subgroups defined by cardinal

<table>
<thead>
<tr>
<th>Model</th>
<th>Subgroups Compared</th>
<th>Simulated Area Between IRFs</th>
<th>MH</th>
<th>Logistic Regression</th>
<th>Raju’s Area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>North vs. South</td>
<td>0.40</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>1PL</td>
<td></td>
<td>0.60</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>East vs. West</td>
<td>0.40</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2PL</td>
<td>North vs. South</td>
<td>0.40</td>
<td>X</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>East vs. West</td>
<td>0.40</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 15. Traditional DIF methods flagging question 1 using a non-proximal approach in CS 2.
direction only are prone to a spatial Simpson’s paradox that has been illustrated here. The counties exhibiting induced regional DIF on question 1 in this case study were located in such a way that DIF was unobservable on this aggregate level. It is of note that this spatial structure is not unrealistic, as the counties exhibiting DIF were selected based upon their inclusion into two major metropolitan statistical areas (MSAs) in Arkansas.

Despite the finding that traditional, non-proximal approaches to regional DIF detection have apparent disadvantages, it was also hypothesized that the local IRT modeling approach proposed would be able to detect this type of regional DIF. Consequently, utilizing the suggested approach to regional DIF detection in the localized IRT modeling framework, Local Moran’s I and associated one-tailed significance testing for spatial clustering was performed on local difficulty estimates (in the 1PL setting) and local discrimination estimates (in the 2PL setting). The results of this significance testing are provided in Figures 32-33 for the 1PL setting and in Figures 34-35 for the 2PL setting. Figure 32 provides the p-values for each of the three different bandwidths at three DIF magnitude levels (small, moderate, and large) using the estimated difficulty parameters obtained from 1PL local IRT model calibrations. Setting again a significance level of 0.05, the p-values ascertained from one-tailed Local Moran’s I significance tests for spatial clustering of the 1PL local IRT model estimated difficulty parameters are mapped in Figure 33, displaying counties in red that have demonstrated significant spatial clustering at all of the three fixed bandwidths selected at each DIF magnitude. Likewise, Figure 34 provides the p-values for each of the three different bandwidths at three DIF magnitude levels using the estimated discrimination parameters obtained from 2PL local IRT model calibrations and Figure 35 further displays these results by denoting counties in red that have demonstrated significant spatial clustering at all of the three fixed bandwidths selected at each DIF magnitude.
Figure 32. 1PL local IRT model local Moran’s I results at three DIF magnitudes (small, moderate, large) for CS 2.
Figure 33. 1PL local IRT model spatial subgroups utilizing local Moran’s I at three DIF magnitudes (small, moderate, large) for CS 2.
Figure 34. 2PL local IRT model local Moran’s I results at three DIF magnitudes (small, moderate, large) for CS 2.
Figure 35. 2PL local IRT model spatial subgroups utilizing local Moran’s I at three DIF magnitudes (small, moderate, large) for CS 2.
Regions identified as having significant spatial clustering can serve as spatial subgroups for use in subsequent, confirmatory DIF analyses. As seen in Figure 33, the Little Rock-North Little Rock-Conway MSA is at least partially identified as a spatial subgroup at all three DIF magnitudes. The Fayetteville-Springdale-Rogers MSA is identified as a spatial subgroup at moderate and large DIF magnitudes. Part of the inability to detect the northwest Arkansas MSA results from a smaller DIF magnitude and, also, the proximity of these counties (Benton, Madison, and Washington) to the border of the study region. However, despite shortcomings, the method does identify spatial subgroups with increasing accuracy as the magnitude of DIF increases, correctly identifying 7 of 9 counties exhibiting potential regional uniform DIF and incorrectly identifying no counties when the DIF magnitude is large. The two counties not identified are Perry and Grant counties, both of which have smaller local sample sizes to begin with and are also located on the outer edges of the latent spatial subgroup. These features of Perry and Grant counties contribute to the misidentification.

As seen in Figure 35, the Little Rock-North Little Rock-Conway MSA is at least partially identified as a spatial subgroup when the non-uniform DIF magnitude is at least moderate. The Fayetteville-Springdale-Rogers MSA is not identified as a spatial subgroup at any DIF magnitude. As in the 1PL local IRT model setting, part of the inability to detect the northwest Arkansas MSA results from the proximity of these counties (Benton, Madison, and Washington) to the border of the study region, facilitating smaller calibration sample sizes. These smaller calibration sample sizes contribute to biased local discrimination estimates at increasing bandwidth sizes (see Figure 31). Sample variability might also have impacted the ability of the method to detect spatial subgroups. While the method is promising, it appears that the 2PL local
IRT modeling approach to regional DIF detection is less capable of accurately identifying non-uniform DIF of any magnitude.

Consider also the circumstance when observed DIF magnitudes are mixed. For instance, one simulation setting was conducted where the most central or populous counties of the MSAs were simulated to have greater DIF magnitudes than the surrounding counties. Consequently, Faulkner, Pulaski, Benton, and Washington counties were simulated such that the area between their respective IRFs and that of counties outside of the MSAs was 0.80. The remaining counties in the MSAs (Grant, Lonoke, Madison, Perry, and Saline counties) were simulated such that the area between their respective IRFs and that of counties outside the MSAs was only 0.40. While the results of local IRT modeling for DIF detection in this circumstance was not a stated objective of the current case study, it is realistic that counties within a spatial subgroup might exhibit varying DIF magnitudes relative to the reference spatial subgroup. The spatial subgroup results for the 1PL local IRT model are provided in Figure 36. It appears as though significant spatial clustering in the mixed DIF magnitude context is most similar to that of a small DIF magnitude setting. This was true also in the 2PL local IRT model setting.

The primary aim of this second case study was to demonstrate the benefit of a localized approach to IRT for regional DIF detection when clustering in the areal units comprising latent spatial subgroups exists. The results for the proposed method utilizing different models (1PL or 2PL) with different DIF types and DIF magnitudes were investigated and compared to traditional, non-proximal DIF detection methods. The current case study offers some evidence that the proposed method can identify regional disparities and latent spatial trends in item functionality that may be unobservable on a more aggregate, global level. However, the current case study also suggests that the utility of the method is limited in terms of identifying non-
uniform regional DIF even when a spatial clustering pattern exists, functioning best in the 1PL local IRT model setting for uniform regional DIF detection instead.

Figure 36. 1PL local IRT model spatial subgroups utilizing local Moran’s I for mixed DIF CS 2.

Case Study 3 – Simulated Data Exhibiting Regional DIF with no Spatial Clustering

This third case study applied both 1PL and 2PL local IRT modeling techniques to Arkansas county data, simulated in a manner similar to the previous case studies, in order to determine if detection of regional DIF was possible when no clear spatial clustering pattern was
present. Data were simulated as in the previous case study with specific factors manipulated such as DIF type and magnitude. As in the two previous simulated case studies, the number of examinees in each county was allocated in accordance with the relative size of the 2016-2017 Arkansas school enrollments, the centroids of each county served as regression points for local model calibration, and three fixed bandwidths of 25 km, 50 km, and 75 km were utilized. However, Arkansas counties were randomly assigned to be either a reference or a focal group member. Of the 75 Arkansas counties, 50 counties were assigned to be reference group members and 25 counties were assigned to be focal group members. This random group assignment insured that one third of the counties exhibited some form of regional DIF but this DIF did not appear to have a strong spatial clustering pattern as can be seen in Figure 37.

Figure 37. Arkansas reference and focal group assignments for CS 3.
Owing to the abundance of case study results, only the most relevant are provided with attention given to those for the first item on the instrument, which was again selected to exhibit regional DIF during the simulation process. The primary aim of this third case study was to determine if there was any utility or benefit to employing the proposed method when a strong spatial clustering pattern does not exist. The functionality of the method in this context was investigated at different levels of DIF type and magnitude, with comparisons made to traditional DIF techniques utilizing a non-proximal approach. The results are presented for the 1PL local IRT and the 2PL local IRT models across all manipulated settings and comparisons are then made to non-proximal DIF detection methods.

**1PL local IRT model results.** The proposed 1PL local IRT method was applied to the Arkansas county 25-item instrument data, simulated to exhibit a small, moderate, or large amount of regional DIF based upon randomized assignment to reference or focal groups (see Figure 37). Tables 16 - 17 provide basic summary statistics for question 1 difficulty estimates and associated standard errors of the 75 local calibrations of the 1PL IRT model for each of the three fixed bandwidths and three DIF magnitudes. Tables 16 - 17 also provide 2PL local IRT model results. The 1PL IRT modeling results for local difficulty estimates for question 1 at each of the three bandwidths and three simulated DIF magnitudes are displayed in Figure 38.

Similar to all previous case studies, the smoothing effect of the bandwidth is evident (see Figure 38 and Tables 16 - 17). In addition, local standard errors are again highest at the 25 km bandwidth for counties with smaller allocated sample sizes. However, these local standard errors decrease with increasing bandwidths and calibration sample sizes. Despite this gradual lowering of average local standard errors across the study region, standard errors for models calibrated along the borders of the study region remain higher relative to more centrally located models.
Table 16. *Local IRT model summary – CS 3 estimated difficulty and discrimination coefficients for question 1 across all 75 Arkansas counties.*

<table>
<thead>
<tr>
<th>Model</th>
<th>Coefficient Estimate Summarized</th>
<th>Simulated Area Between IRFs</th>
<th>25 km Bandwidth</th>
<th>50 km Bandwidth</th>
<th>75 km Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Min.</td>
<td>Max.</td>
<td>Mean</td>
</tr>
<tr>
<td>1PL</td>
<td>Difficulty</td>
<td>0.40</td>
<td>0.099</td>
<td>-0.338</td>
<td>0.623</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>0.169</td>
<td>-0.338</td>
<td>0.905</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>0.223</td>
<td>-0.338</td>
<td>0.988</td>
</tr>
<tr>
<td>2PL</td>
<td>Discrimination</td>
<td>0.40</td>
<td>0.926</td>
<td>0.582</td>
<td>1.381</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>0.883</td>
<td>0.310</td>
<td>1.381</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>0.881</td>
<td>0.303</td>
<td>1.381</td>
</tr>
</tbody>
</table>
Table 17. *Local IRT model summary – CS 3 estimated difficulty and discrimination coefficients for question 1 across reference and focal Arkansas counties.*

<table>
<thead>
<tr>
<th>Model</th>
<th>Coefficient Estimate Summarized</th>
<th>Simulated Area Between IRFs</th>
<th>25 km Bandwidth</th>
<th>50 km Bandwidth</th>
<th>75 km Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Min.</td>
<td>Max.</td>
<td>Mean</td>
</tr>
<tr>
<td>1PL</td>
<td>Reference Group Difficulty</td>
<td>0.40</td>
<td>-0.043</td>
<td>-0.338</td>
<td>0.190</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>-0.043</td>
<td>-0.338</td>
<td>0.190</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>-0.043</td>
<td>-0.338</td>
<td>0.190</td>
</tr>
<tr>
<td></td>
<td>Focal Group Difficulty</td>
<td>0.40</td>
<td>0.384</td>
<td>0.172</td>
<td>0.623</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>0.592</td>
<td>0.381</td>
<td>0.905</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>0.753</td>
<td>0.562</td>
<td>0.988</td>
</tr>
<tr>
<td>2PL</td>
<td>Reference Group Discrimination</td>
<td>0.40</td>
<td>1.001</td>
<td>0.726</td>
<td>1.381</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>1.001</td>
<td>0.726</td>
<td>1.381</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>1.001</td>
<td>0.726</td>
<td>1.381</td>
</tr>
<tr>
<td></td>
<td>Focal Group Discrimination</td>
<td>0.40</td>
<td>0.777</td>
<td>0.582</td>
<td>0.966</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>0.647</td>
<td>0.310</td>
<td>0.804</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>0.641</td>
<td>0.303</td>
<td>1.010</td>
</tr>
</tbody>
</table>
The Arkansas county data for the current simulated case study, despite including 25 focal group member counties that should significantly differ from the 50 reference group member counties in terms of question 1 item functionality, was not created with a highly discernible spatial clustering pattern. Consequently, the 25 km bandwidth which is most equivalent to areal unit calibrations of the global 1PL IRT model most clearly display regional differences in local

*Figure 38.* CS 3 1PL local IRT model difficulty estimates across fixed bandwidths and three DIF magnitudes (small, moderate, large).
difficulty estimates, especially at higher DIF magnitudes (see Figure 38). However, as the bandwidth increases to 50 km and 75 km, only the greatest clustering of these focal group counties in the southwest corner of Arkansas emerges as a possible spatial subgroup. While this might suggest the existence of a spatial subgroup, the region visually exhibits only a very weak nonstationarity in local difficulty estimates even at the largest DIF magnitude (see Figure 38).

Based upon these 1PL local IRT modeling results, the proposed method utilizing several fixed bandwidths to smooth the local difficulty estimate surface may struggle to correctly identify areal units that are part of a focal group if focal group members do not have a strong spatial clustering pattern. This is to be expected though as local spatial modeling techniques necessarily assume a significant level of spatial nonstationarity. However, despite this potential disadvantage to the proposed method, there is still value in visualizing IRT difficulty estimates across space. This visualization, made possible by the localization of IRT modeling, allows for an analysis of the sensitivity and/or stability of the global model parameter estimates, facilitates spatial interpretation of findings, and can suggest subsequent DIF analyses (Fotheringham et al., 2002; Matthews & Yang, 2012).

**2PL local IRT model results.** The 2PL model of the proposed method for regional DIF detection was also applied to simulated case study data. Tables 16 - 17, previously presented, provide basic summary statistics for question 1 discrimination estimates and associated standard errors of the 75 local calibrations of the 2PL IRT model for each of the three fixed bandwidths and three DIF magnitudes. Surface mappings of county-specific discrimination estimates for question 1 of the instrument are provided in Figure 39.
Figure 39. CS 3 2PL local IRT model discrimination estimates across fixed bandwidths and DIF magnitudes (small, moderate, large).

The smoothing feature of the bandwidth is again evident in Figure 39. While similar to the 1PL local IRT model setting in that the clearest display of simulated regional differences in local coefficient estimates occurs at the 25 km bandwidth, especially at higher DIF magnitudes, the clarity of these differences visually is lessened when utilizing a 2PL local IRT modeling approach and when non-uniform regional DIF is present. While group distinctions are less
apparent visually in this non-uniform DIF setting, larger bandwidths do facilitate the visual identification of a potential spatial subgroup in the southwest corner of Arkansas at higher DIF magnitude levels. As in prior case study findings, this lack of visual evidence for regional DIF could be attributed to sampling variability, graphical decisions such as the fill scale utilized and the color palette applied, or an inherent difficulty in identification of non-uniform regional DIF. However, it appears that the most influential factor in the lack of correctly identifiable focal group membership in the current case study is the lack of a spatial clustering pattern. The proposed methodology assumes a spatial structure in order to be most beneficial. The ability of the 2PL local IRT model to detect non-uniform regional DIF will be discussed further with comparisons made to traditional, non-proximal DIF detection techniques.

**Regional DIF detection.** The current case study was presented to explore the relative advantages and disadvantages of local IRT modeling for regional DIF detection when a latent spatial structure, specifically a clustering pattern, is not definitively present. Rather, the spatial structure of focal group members in a region might be of a form where multiple group DIF detection techniques, many pairwise comparisons, or even non-proximal approaches based upon cardinal direction might be more appropriate. Unlike the prior case studies, it was hypothesized that the absence of a spatial clustering pattern in the current case study would make detection of regional DIF challenging for the proposed method, which attempts to identify spatial subgroups for subsequent analyses utilizing local indicators of spatial association.

To investigate this hypothesis, Mantel-Haenszel (MH) chi-square, logistic regression, and Raju’s EUA methods were applied to the various simulated datasets, comparing subgroups North to South and East to West for each model type and DIF magnitude setting. These three methods were used to test for both uniform and non-uniform DIF on question 1 of the instrument at a
Table 18. Traditional DIF methods flagging question 1 using a non-proximal approach in CS 3.

<table>
<thead>
<tr>
<th>Model</th>
<th>Subgroups Compared</th>
<th>Simulated Area Between IRFs</th>
<th>MH</th>
<th>Logistic Regression</th>
<th>Raju’s Area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>North vs. South</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>1PL</td>
<td></td>
<td>0.40</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>East vs. West</td>
<td>0.40</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>North vs. South</td>
<td>0.40</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2PL</td>
<td></td>
<td>0.60</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>East vs. West</td>
<td>0.40</td>
<td>X</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.80</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

For the six different DIF settings in this case study, two pairwise comparisons involving cardinal direction were assessed with three different methods. In the 12 pairwise comparisons possible, MH identified DIF in East to West comparisons for all magnitudes of uniform DIF in the 1PL setting and, also, identified regional DIF in North to South comparisons at the largest DIF magnitude simulated. However, the MH method was unable to identify any of the non-uniform DIF simulated in the 2PL setting, regardless of magnitude. Logistic regression

significance level of 0.05. Table 18 is provided to summarize these non-proximal, pairwise comparisons where ‘X’ indicates the first instrument item was not flagged for DIF and ‘✓’ indicates the first instrument item was flagged for DIF by the method.
performed only marginally better, again identifying all of the East to West uniform DIF present in the 1PL setting but also identifying the East to West non-uniform DIF present in the 2PL setting at the largest DIF magnitude. Raju’s EUA method performed the best of the three traditional methods identifying uniform DIF at a comparable rate to the other two methods but identifying non-uniform DIF more frequently. However, this superior performance of Raju’s EUA method would not be surprising given that DIF magnitude was manipulated by altering the area between the reference and focal group member IRFs. In addition, this apparent ability to identify non-uniform DIF in this setting with Raju’s EUA method comes with the increased potential for type I errors that was seen in the first case study. While the relative merits of the three methods will not be discussed here, it does appear that non-proximal approaches can detect the presence of DIF on question 1 in the current case study setting though results will depend upon the proportion of randomly distributed DIF counties located within each proximal region.

Despite the finding that traditional, non-proximal approaches to regional DIF detection have apparent advantages and are in fact capable of identifying the presence of DIF at a regional level defined by cardinal direction, the approaches do not go further unless data is recursively partitioned and disaggregated at arbitrary boundaries defined once again by cardinal direction or some other means. The non-proximal approaches do not in and of themselves identify which counties are producing the significant DIF findings for question 1. In addition, these findings still might result in a spatial Simpson’s paradox. In fact, mappings of estimated difficulty and estimated discrimination parameters in Figures 38 and 39, respectively, indicated that not all of the eastern counties differed from the western counties in item parameter estimates and potential item functionality on question 1 yet this is not suggested directly by the results of the non-proximal approaches to DIF detection. Given this consideration, it was also hypothesized that the
local IRT modeling approach proposed would serve as a microscope, uncovering further spatial patterns and suggesting subsequent analyses that a non-proximal approach would not.

For comparative purposes, the suggested approach to regional DIF detection in the localized IRT modeling framework utilizing Local Moran’s I and associated one-tailed significance testing for spatial clustering was performed on local difficulty estimates (in the 1PL setting) and local discrimination estimates (in the 2PL setting). The results of this significance testing are provided in Figures 40-41 for the 1PL setting and in Figures 42-43 for the 2PL setting. Figure 40 provides the p-values for each of the three different bandwidths at three DIF magnitude levels (small, moderate, and large) using the estimated difficulty parameters obtained from 1PL local IRT model calibrations. Setting again a significance level of 0.05, the p-values ascertained from one-tailed Local Moran’s I significance tests for spatial clustering of the 1PL local IRT model estimated difficulty parameters are mapped in Figure 41, displaying counties in red that have demonstrated significant spatial clustering at all of the three fixed bandwidths selected at each DIF magnitude. Likewise, Figure 42 provides the p-values for each of the three different bandwidths at three DIF magnitude levels using the estimated discrimination parameters obtained from 2PL local IRT model calibrations. Figure 43 further displays these results by denoting counties in red that have demonstrated significant spatial clustering at all of the three fixed bandwidths selected at each DIF magnitude.

Regions identified as having significant spatial clustering can serve as spatial subgroups for use in subsequent, confirmatory DIF analyses. As seen in Figure 41, there are no large or clearly discernible spatial subgroups identified at all three DIF magnitudes in the 1PL setting. However, Figure 40 complements the previous mappings of estimated difficulty parameters from the 1PL local IRT models at three fixed bandwidths for the various DIF magnitudes in revealing
Figure 40. 1PL local IRT model local Moran’s I results at three DIF magnitudes (small, moderate, large) for CS 3.

both individual counties and specific regions in the state that might be contributing to the
significant DIF findings in the non-proximal approach methods. For instance, it appears that
counties affecting the results of DIF comparisons East to West are mostly located in the central
east and southwest regions of the state, rather than in the north. Moreover, if these results are
Figure 41. 1PL local IRT model spatial subgroups utilizing local Moran’s I at three DIF magnitudes (small, moderate, large) for CS 3.
Figure 42. 2PL local IRT model local Moran’s I results at three DIF magnitudes (small, moderate, large) for CS 3.
Figure 43. 2PL local IRT model spatial subgroups utilizing local Moran’s I at three DIF magnitudes (small, moderate, large) for CS 3.
are also compared to the surface mappings of the estimated difficulty parameters from the 1PL local IRT models (see Figure 38), the locations of counties with focal group membership become more apparent and additional visual information is provided regarding any presumed spatial patterns in the state that would not be possible in the non-proximal approach.

The same results are discovered in the 2PL setting of this case study, with no large or clearly discernible spatial subgroups identified at any of the three DIF magnitudes (see Figure 43). However, once more, visualization of significant spatial clustering p-values (see Figure 42) serves as a supplement to the previous mappings of estimated discrimination parameters from the 2PL local IRT models at three fixed bandwidths for the various DIF magnitudes (see Figure 39). Both visualizations help uncover further spatial patterns and suggest subsequent analyses that a non-proximal approach would not such as the investigation of a specific county or of a further disaggregated set of spatial subgroups defined by cardinal direction (e.g., northwest, central west, southwest, northeast, central east, and southeast).

The primary aim of this third case study was to determine if there was utility or benefit in a localized approach to IRT for regional DIF detection when strong spatial clustering in the areal units does not exist. The results for the 1PL local IRT and the 2PL local IRT models across different DIF types and DIF magnitudes were investigated and compared to traditional, non-proximal DIF detection methods. The current case study offers some evidence that, while the localized approach to IRT for regional DIF detection may struggle to correctly identify areal units that are part of a focal group if focal group members do not have a strong spatial clustering pattern, there is still value in the visualization afforded by the proposed method. This visualization may still help to identify regional disparities and latent spatial trends in item functionality that may be otherwise unobservable using approaches that do not account for the
proximity of areal units in a study region. It should be noted, however, that the current case study once more suggests that the utility of the method is most limited in the identification of non-uniform regional DIF.

**Case Study 4 – Empirical Data Application**

Differences in item functionality are of concern across a broad range of disciplines where the use of survey instruments to measure latent traits is essential. While the third simulated case study attempted to closely imitate an empirical study, the application of the proposed methodology to actual survey data gathered from the 2015-16 Malawi DHS will allow for an illustration of the localized approach to IRT for regional DIF detection when realistic issues such as non-regularity of regression points and true differences in local calibration sample sizes are present. The 1PL local IRT modeling approach with model calibration at 28 district centroids, using four fixed bandwidths will be applied in this empirical data case study with the goal of identifying potential regional DIF in a 13 item instrument measuring family planning method knowledge taken from the most recent DHS data collected in Malawi.

Malawi is a poor, rural country in Sub-Saharan Africa and while knowledge of at least one contraceptive method is now nearly universal according to the NSO (2017), there is still a substantial unmet need for family planning. According to the Guttmacher Institute (2014), 54% of pregnancies in Malawi were unintended with approximately one third of those unintentional pregnancies leading to abortion and another one third leading to miscarriage. The infant mortality rate in Malawi is also high, a majority of women still do not obtain adequate prenatal and delivery care, and maternal mortality rates remain elevated as many women face high-risk pregnancies owing to their age (either young or old), the close spacing of their pregnancies,
and/or the sheer number of their pregnancies (Guttmacher Institute, 2014). Consequently, there is a benefit to meeting the family planning need of Malawian women.

While likely reasons for the current unmet contraceptive need in Malawi include limited access, fear of side-effects, and poor quality of services, other reasons include limited choice, gender-based barriers, cultural and religious opposition, and provider bias (WHO, 2016). In addition, one of the most fundamental reasons for unmet contraceptive need and unintended pregnancies in Malawi stems from a lack of family planning method knowledge with men and women sometimes employing “guesswork” before education about contraceptive methods (Shattuck et al., 2011). Education by fieldworkers and healthcare facilities regarding contraceptive methods could, therefore, increase the knowledge of all contraceptive methods and allow women to make more informed and effective contraceptive choices even following discontinuation of a previously used method, limiting the risks of unintended pregnancy to mother, child, and family. Regional differences in the knowledge of particular family planning methods can help guide policy makers and steer educational interventions appropriately.

The 13 survey items intended to measure women’s knowledge of family planning methods (see Table 3) might function differently across Malawi due to geographic location. The primary aim of this empirical case study was to demonstrate the localized approach to IRT for regional DIF detection in context with comparisons made to non-proximal DIF techniques based upon pairwise comparisons of the Malawian regions (Northern, Central, Southern).

Local IRT model specifications. The local IRT model selected for use in the current empirical case study was a 1PL local IRT model with model calibration at 28 Malawi district centroids using four fixed bandwidths. The rationale behind regression points being allocated to the 28 district centroids was discussed previously and was meant to avoid bias attributable to the
geographic off-setting of GPS data in the DHS program. The four fixed bandwidths were selected at various, increasing distances to create additional connectedness in the neighborhood structures that were complicated by the irregularity in the Malawi district centroid locations. The neighborhood structures that are formed by these four specified bandwidths can be seen in Figure 44. The smallest bandwidth \((h = 20\text{km})\) is equivalent to an areal unit calibration of the global model. The 80 km bandwidth, however, results in a neighborhood structure that has an average of 3.4 links for each calibration point. At this distance, due to the irregularity in the spatial location of the district centroids, seven of the districts have only one link while Blantyre has 9 links. The spatial neighborhood formed by utilizing an 80 km bandwidth highlights the non-regularity in the district areal units, with districts in the southern and central regions of Malawi much more connected than districts in the northern region of Malawi. The 140 km bandwidth expands the spatial neighborhoods, having an average of 7.9 links. However, at 140 km, two districts (Chitipa and Karonga) still have only 2 links while Balaka now has 13 links. As the

![Figure 44. Neighborhood structures for Malawi districts at four fixed bandwidths.](image)
bandwidth is increased to 200 km, calibration points now have an average of 11.4 links but Chitipa still has only 2 links while Ntcheu has 17. According to Fotheringham et al. (2002, p. 46), use of fixed bandwidths in the presence of this non-regularity of calibration points creates the potential for undersmoothed surfaces of parameter estimates. The use of several fixed bandwidths in the current case study will allow for the further investigation of the previously observed bandwidth smoothing properties in a less ideal setting.

The 1PL local IRT model was selected over the 2PL local IRT model for two primary reasons, based upon computational speed and hypothesized DIF results. First, while variability in machines can affect the computational efficiency of the proposed method as well as characteristics of the specific analysis such as global and local sample sizes, instrument length, number of local calibrations utilized, and bandwidth size, the results of timing local IRT modeling in the previous case studies suggested that use of the 2PL local IRT model takes approximately 1.8 to 2.0 times longer than the 1PL local IRT model in terms of both user and elapsed time. Consequently, any exploratory data analysis utilizing this method should first be conducted with the 1PL local IRT model. Second, owing to the nearly universal knowledge of contraception in Malawi, uniform regional DIF demonstrating that the probability of knowing a particular family method was uniformly higher or lower across the knowledge continuum for specific districts or locations in Malawi was of more interest and appeared more promising for subsequent regional investigation, analysis, or intervention. Consequently, only a 1PL local IRT model was applied to the 13 item survey data.

Unidimensionality assumption. The 13 item dichotomously scored instrument in the current case study is intended to measure one latent trait, family planning method knowledge. Unidimensionality is an assumption of the proposed method since the local IRT models proposed
are unidimensional IRT models. Therefore, several approaches to select the number of factors to retain in a factor analysis of the data were utilized including Kaiser’s rule, scree test, optimal coordinate (OC), acceleration factor (AF) and parallel analysis (PA) methods. While most methods suggested two factors be retained, the scree test and AF methods suggest retaining only one factor. Additionally, the first eigenvalue of 5.12 is approximately three times larger than the next eigenvalue of 1.61 and is much greater than all other eigenvalues, suggesting that the instrument approximates an essentially unidimensional set of items.

A one-factor oblique principal-axis factor analysis was conducted accounting for 35% of the total variability in the set of items. All factor loadings for items ranged between 0.40 and 0.74, which meets a 0.40 criterion. Cronbach’s coefficient alpha, a measure of internal consistency reliability, was approximately 0.84 for the full 13 item instrument and corrected values range from 0.83 to 0.84. While observing sufficient internal consistency is not a true measure of unidimensionality, it does provide additional support for the use of the 13 item instrument as an approximately unidimensional set of items.

Moreover, as noted previously, unidimensionality is desirable for scale construction (Hambleton & Swaminathan, 1985, p. 156) but violations of unidimensionality may or may not be problematic (de Ayala, 2009, p. 20). As the current argument for unidimensionality is being made to investigate items for DIF, rather than for scale construction purposes, a unidimensional IRT model seems reasonable for this application and will be subsequently assumed.

**1PL local IRT model results.** The 1PL model of the proposed method for regional DIF detection was applied to the DHS 13 item family planning method knowledge instrument data. Table 19 provides the 1PL local IRT difficulty estimates and associated standard errors for all 13 items averaged across the 28 local calibrations of the 1PL model for each of the four fixed
Table 19. Local IRT model summary – CS 4 estimated difficulty coefficients for DHS items averaged across all 28 Malawi districts.

<table>
<thead>
<tr>
<th>Item</th>
<th>20 km Bandwidth</th>
<th>80 km Bandwidth</th>
<th>140 km Bandwidth</th>
<th>200 km Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Min.</td>
<td>Max.</td>
<td>Mean SE</td>
</tr>
<tr>
<td>V304$01 (Q1)</td>
<td>-3.366</td>
<td>-4.213</td>
<td>-2.658</td>
<td>0.160</td>
</tr>
<tr>
<td>V304$02 (Q2)</td>
<td>-2.138</td>
<td>-2.814</td>
<td>-1.446</td>
<td>0.113</td>
</tr>
<tr>
<td>V304$03 (Q3)</td>
<td>-3.893</td>
<td>-4.871</td>
<td>-3.106</td>
<td>0.194</td>
</tr>
<tr>
<td>V304$05 (Q4)</td>
<td>-4.274</td>
<td>-5.378</td>
<td>-3.549</td>
<td>0.221</td>
</tr>
<tr>
<td>V304$06 (Q5)</td>
<td>-2.609</td>
<td>-3.416</td>
<td>-1.847</td>
<td>0.128</td>
</tr>
<tr>
<td>V304$07 (Q6)</td>
<td>-0.837</td>
<td>-1.885</td>
<td>-0.321</td>
<td>0.091</td>
</tr>
<tr>
<td>V304$08 (Q7)</td>
<td>-0.696</td>
<td>-1.816</td>
<td>0.382</td>
<td>0.090</td>
</tr>
<tr>
<td>V304$09 (Q8)</td>
<td>-1.127</td>
<td>-2.197</td>
<td>-0.148</td>
<td>0.095</td>
</tr>
<tr>
<td>V304$11 (Q9)</td>
<td>-3.093</td>
<td>-4.109</td>
<td>-2.214</td>
<td>0.147</td>
</tr>
<tr>
<td>V304$13 (Q10)</td>
<td>-1.307</td>
<td>-2.320</td>
<td>-0.395</td>
<td>0.097</td>
</tr>
<tr>
<td>V304$16 (Q12)</td>
<td>0.288</td>
<td>0.0382</td>
<td>0.818</td>
<td>0.086</td>
</tr>
<tr>
<td>V304$18 (Q13)</td>
<td>0.315</td>
<td>-0.532</td>
<td>0.949</td>
<td>0.086</td>
</tr>
</tbody>
</table>
bandwidths. As the discrimination parameter was constrained to one for all model calibrations, this value is excluded from the table.

Despite the non-regular spacing of the Malawi district centroids, which served as regression points for local model calibration, the smoothing effect of the bandwidth is still observed in Table 19. This smoothing effect, whereby the range and variability in the local parameter estimates reduces with increasing bandwidth sizes that directly produce increasing local sample sizes, is apparent in this empirical case study also. The greatest smoothing of local difficulty estimates occurs as the bandwidth increases from 20 km to 80 km, which can be attributed to the fact that the 20 km bandwidth is essentially an areal unit calibration of IRT models whereas the 80 km bandwidth has a true spatial neighborhood structure linking districts to one another in local model calibrations. Smoothing of the local difficulty estimates also occurs when the bandwidth size is increased from 80 km to 140 km. However, this smoothing appears to slow, with local difficulty estimates stabilizing, as the bandwidth size is increased from 140 km to 200 km. This is presumably due to the fact that the districts located in the central and southern regions of Malawi are very connected at both bandwidths but the more northern districts of such as Chitipa and Karonga are still fairly disconnected with survey data from very few neighboring districts incorporated into these local model calibrations. The smoothing effect of the bandwidth will be better visualized with mappings of the local difficulty estimates. Local standard errors do decrease as the bandwidth increases, with standard errors at smaller bandwidths sizes being most related to district sample sizes while standard errors at larger bandwidths being more related to the proximity of the district to other districts and to the edge of the spatial region.
It is notable that the difficulty estimates when averaged across all 28 local model calibrations are almost all negative values across all of the four bandwidths. This suggests that the items on this survey are very ‘easy’ but, also, that they are most discriminating among respondents of lower family planning method knowledge. The fourth item is the easiest, which is not surprising as this is the item used to assess respondent knowledge of the male condom and nearly 97% of all DHS survey respondents stated their knowledge of the male condom as a contraceptive method. According to the Guttmacher Institute (2014), the injectable is the most commonly used modern and non-permanent method of contraception in Malawi followed by the pill. Moreover, condom usage (both male and female) has been widely advertised as an effective form of contraception that provides HIV protection (John, Babalola, & Chipeta, 2015). Consequently, it is also not surprising that items 1, 3, and 11 representing knowledge of the pill, injections, and the female condom respectively are also very easy items with between 92% and 95% of respondents stating their knowledge of the methods for family planning purposes.

Overall, the people of Malawi do have knowledge of family planning methods and this knowledge is not limited to certain districts. Assuming that the total survey score for the 13 items is a representation of respondent knowledge of family planning methods, it is apparent that the people in Malawi are knowledgeable with an overall total score mean of approximately 10.05 and a standard deviation of approximately 2.90. The district level total scores are all very similar across Malawi with district averages ranging between 9.42 and 10.87. These district level averages and the district level standard deviations for total scores are provided in Figure 45 below. Both the average total scores for the 13 item survey instrument and their corresponding standard deviations at the district level are all fairly similar across the country.
Despite the widespread knowledge in Malawi of these methods, partner dynamics and sexual pleasure seeking can prevent use of well-known contraceptive methods (John et al., 2015), condom use can be inconsistent and is generally only effective for contraception when HIV or other infections are present, and method adherence for injections and the pill is low (Dasgupta, Zaba, & Crampin, 2015). In fact, Dasgupta et al. (2015) found that only 28% of women using the pill obtained their prescription refill on time and only 15% of women using the injection method continued for a full year without experiencing a gap between injections that would put them at risk for unintended pregnancy. In order to investigate regional differences in knowledge of family planning methods that interfere least with partner dynamics and that
necessitate correct and consistent use, the rest of this case study presents local IRT model results and regional DIF analyses for only two survey items regarding IUD and implant knowledge (questions 2 and 9). Surface mappings of 1PL local IRT results, local Moran’s I results, and suggested spatial subgroups are provided in Figures 46 - 48 for the IUD item and in Figures 49 - 51 for the implant item.

**Regional DIF detection.** The current case study was conducted to determine if any of the Malawi DHS 13 survey items intended to measure family planning method knowledge exhibited regional DIF with comparisons made to non-proximal DIF techniques based upon pairwise comparisons of the Malawian regions (Northern, Central, Southern).

Malawi is divided into three political regions (see Figure 22). Mantel-Haenszel chi-square, logistic regression, and Raju’s EUA methods were applied to detect item-level DIF for the 13 survey items in all regional pairwise comparisons (i.e., North to South, North to Central, and Central to South) at a significance level of 0.05. Several items were flagged for DIF, especially when comparison was made between the North and Central regions. Most notable, however, was the flagging of questions 2 and 9 for DIF using all three non-proximal approaches when pairwise comparisons were made between certain regions. Recall that these two items will be the items of primary interest for the current case study. DIF was detected for question 2, assessing respondent knowledge of IUDs, when comparisons were made North to South and North to Central. DIF was detected for question 9, assessing respondent knowledge of implants, when comparisons were made North to Central and South to Central.

Despite finding that these items had been flagged as exhibiting potential DIF, further information regarding the individual districts (rather than the more aggregate regions) was not directly provided using these non-proximal techniques unless all pairwise comparisons among
the 28 districts were to be conducted. However, treating local IRT modeling as a spatial microscope for the investigation of regional item-level DIF, surface mappings of district difficulty estimates for the IUD item are provided in Figure 46. These mappings are provided along with accompanying maps of local Moran’s I values (see Figure 47) and suggested spatial subgroups of districts based upon one-tailed significance testing for spatial clustering of local difficulty estimates (see Figure 48). Setting the significance level at 0.05, Figure 48 displays districts in red that have demonstrated significant spatial clustering at all of the four fixed bandwidths selected as well as districts that have demonstrated significant spatial clustering at all bandwidths beyond the 20 km specification. It appears that women in the three northernmost districts of Chitipa, Karonga, and Rumphi have a lower probability of knowing the IUD contraceptive method at the same level of family planning method knowledge than women in more central and southern districts. While potentially due to an undersmoothing of local difficulty estimates in these northern districts, this might suggest that not all districts in the northern region have the same level of knowledge regarding the IUD despite having similar family planning method knowledge. These regional differences may be due in part to educational programs regarding contraception that have been working in the Karonga district of Malawi or other factors such as the northern region being the most rural region of Malawi but also having the highest education levels, highest literacy rates, and greatest number of polygamous relationships (Baschieri et al., 2013). A complex interaction of these factors and others might be contributing to the regional differences in item functionality of the IUD knowledge survey item and further investigation would be warranted, especially as the IUD could be a very promising family planning method for women in these Malawi districts.
Figure 46. CS 4 1PL local IRT model difficulty estimates across fixed bandwidths for the DHS IUD item.
Figure 47. CS 4 1PL local IRT model local Moran’s I results across fixed bandwidths for the DHS IUD item.
Figure 48. 1PL local IRT model spatial subgroups utilizing local Moran’s I at four fixed bandwidths for the DHS IUD item.
Surface mappings of district difficulty estimates for the implant survey item are provided in Figure 49. These mappings are also provided along with accompanying maps of local Moran’s I values (see Figure 50) and suggested spatial subgroups of districts based upon one-tailed significance testing for spatial clustering of local difficulty estimates (see Figure 51). It appears that women in the central districts as well as in the district of Zomba have a higher probability of knowing the implant contraceptive method at the same level of family planning method knowledge than women in northern and southern districts. These regional differences may be due in part to the proximity of districts in the central region to urban centers with greater potential for family planning method advising than in the more northern districts or in the southern districts. However, more investigation would need conducted to ascertain the reasons behind these apparent regional differences in item functionality.

While the non-regular spacing of the Malawi district centroids that served as regression points may have resulted in undersmoothing of local difficulty estimates in local IRT model calibrations, the method did allow for visualization and further interpretation of aggregate DIF results ascertained from non-proximal methods. For instance, consider that DIF detection utilizing traditional pairwise comparison techniques could also be applied to look for differences in item functionality between Malawi rural and urban areas as classified in Figure 52. Whether comparisons are made utilizing Malawi regions (North to Central, North to South, and South to Central) or made utilizing district classification as rural or urban, any detected differences in item functionality would still not be able to reveal the magnitude, the scope, or the spatial trend of those regional disparities. Figure 53 provides item characteristic curves for the IUD and implant survey items at various levels of data aggregation, demonstrating the unobservable magnitude of DIF at the more disaggregate district level when comparison is made to the
Figure 49. CS 4 1PL local IRT model difficulty estimates across fixed bandwidths for the DHS implant item.
Figure 50. CS 4 1PL local IRT model local Moran’s I results across fixed bandwidths for the DHS implant item.
Figure 51. 1PL local IRT model spatial subgroups utilizing local Moran’s I at four fixed bandwidths for the DHS implant item.
magnitude of DIF in the items observed at more aggregate levels of region or rural/urban classification. While item characteristic curves such as those provided in Figure 53 help to put into perspective the magnitude of differences in item functionality that exist between areal units, the visualization is still not as enlightening as the mapped local IRT modeling results that take into consideration with their visualization both the proximity and geographic location of districts.

![Malawi Urban and Rural Districts](image)

*Figure 52. Malawi urban and rural district classifications for CS 4.*

The current case study of 13 Malawi DHS survey items intended to measure knowledge of family planning methods was conducted to demonstrate the use of the proposed local IRT model for regional DIF detection in context with comparisons made to non-proximal DIF techniques. Moreover, the empirical case study demonstrated the use of the proposed method with fixed bandwidths on a very irregular spatial lattice. While greater familiarity with Malawian family planning practices would be necessary to suggest subsequent follow-up analyses, there
was still value in visualizing the spatial patterns of and regional differences in specific family planning method knowledge throughout Malawi. Additionally, there was value in the application of the proposed local IRT modeling approach to empirical data.

The proposed method serves to detect DIF across space without a priori groupings, identifying regional disparities and latent spatial trends that may otherwise be unobservable. Results for three simulated case studies and one empirical case study were provided to illustrate the use of the proposed method. Simulated case studies provided a demonstration of the method under the manipulation of several factors that could not be controlled empirically such as DIF type and magnitude while the empirical case study provided a demonstration of the method in a realistic context where additional complexities in local modeling emerge such as unequal calibration sample sizes and irregular spacing of model calibration points.
Figure 53. Item characteristic curves for the DHS IUD and implant items by region, urban classification, and district.
Chapter 5
Discussion

The purpose of the current study was both to propose a method to detect item-level regional DIF and to illustrate the proposed method by providing applications in the form of several case studies, with comparisons made to traditional DIF techniques. Methodological choices such as model type and bandwidth size were explored. Additionally, characteristics of items exhibiting regional DIF were manipulated during data simulation such as the magnitude and nature of DIF, the spatial arrangement of groups exhibiting DIF, and the local sample sizes in these groups in order to determine the advantages and disadvantages of the proposed method under different circumstances. The localized approach to IRT modeling for regional DIF detection as well as the simulated and empirical case study results will be discussed, with mention both of the current literature and of the additions this research makes to the literature. Lastly, based upon the findings of this study, implications and suggestions for regional DIF detection, limitations of the current research, and future research directions will be presented.

Summary of Local Item Response Theory

The proposed localized approach to IRT modeling and DIF detection provided a flexible framework wherein 1PL or 2PL IRT models were fit using MMLE, with local calibrations at select regression points utilizing subsets of global data which were incorporated into each local model based upon a defined spatial neighborhood structure and a box-car kernel weighting scheme that allowed the models to essentially sweep across a study region in a moving window technique producing spatially-varying IRT parameter estimates. Using the idea that violations of parameter invariance across identifiable spatial or regional subgroups provided a working definition of regional DIF in space, mappings of these spatially-varying local IRT item parameter
estimates allowed for visual investigation of potential DIF (both uniform and non-uniform) based upon geographical location without need for pre-specified groupings and before any confirmatory DIF testing. The subsequent identification of latent spatial subgroups could be accomplished by treating IRT item parameter estimates as fixed values that could be assessed for spatial clustering utilizing a local indicator of spatial association such as Local Moran’s I and corresponding significance testing procedures, providing further support to the regional DIF detection procedure proposed. Owing to the location-dependent nature of the proposed methodology, all results were mapped for visualization of spatial patterns.

Investigation of DIF throughout a spatial region has typically relied on non-proximal approaches whereby pairwise comparisons based upon cardinal direction or all possible pairwise comparisons based upon arbitrary spatial boundaries are made using traditional DIF techniques. Even when the number of pairwise comparisons is reduced by methods that essentially amount to artificially creating two groups, one group against the aggregate of the others or against the composite of all, this partitioning of the multiple groups to adhere to DIF techniques designed for comparing only two groups (i.e., reference and focal) is uninformed. All of the above stated procedures partition the data consisting of multiple groups into two non-informative, arbitrary groups. While some techniques have developed to extend DIF methodology to the multiple group setting (e.g., Berger & Tutz, 2016; Bock & Zimowski, 1997; Finch, 2005; Magis, Tuerlinckx & De Boeck, 2015; Muthen & Asparouhov, 2014; Strobl, Kopf & Zeileis, 2010; Tutz & Berger, 2016; Woods & Grimm, 2011), none of these techniques has yet to take into consideration the proximity of the groups to one another in space. The spatial structure of the groups is not fully utilized or is dismissed in the current literature.
The local IRT model for regional DIF detection proposed in this work adds to the current literature by accounting for spatial structure in these multiple groups. The proposed method addresses nonstationarity directly by allowing parameter estimates to vary across space, considers spatial structure and proximity by incorporating a neighborhood structure with a box-car kernel weighting scheme in local model calibrations, and provides a visualization of relationships and spatial trends that exist among the many groups of interest in a study region. This proposed method adds a spatial analytic technique to the investigation of DIF and, owing to the location dependent results of the localized approach, also adds a visualization tool to the literature that can reveal underlying spatial patterns, facilitate interpretation of regional DIF findings, and suggest subsequent analysis.

**Summary of Simulated Case Study Results**

The advantages and the disadvantages of the proposed method were explored using three simulated case studies. Each case study was designed to illustrate hypothesized benefits and potential drawbacks of the localized approach to IRT modeling for regional DIF detection in a variety of settings, with comparisons made to non-proximal DIF techniques. Factors such as DIF type and magnitude as well as spatial structure were manipulated while studying the sensitivities of the method to procedural choices such as model choice and bandwidth size.

The first case study demonstrated the benefit of the proposed method when no regional DIF is present. Non-proximal approaches to DIF detection utilizing pairwise comparisons based upon cardinal direction correctly indicated that no regional DIF was present. However, the potential for making an incorrect conclusion of no DIF at this more aggregate level, essentially amounting to a spatial Simpson’s paradox, was very possible with this traditional approach to DIF detection. Mappings of estimated local IRT model parameters allowed for a visualization of
any potential regional DIF that would mitigate these concerns, demonstrating value in the proposed method. Additionally, when comparisons were made to a type of non-proximal DIF detection technique where each areal unit subgroup is compared to the aggregate of the remaining areal subgroups, the proposed method yielded comparable results in terms of type I errors. The first case study provided evidence that the proposed method would yield similar results to traditional methods when no regional DIF was present but could provide additional visual assurance of any non-significant DIF findings.

The second case study demonstrated the benefit of the proposed method when regional DIF is present and a spatial clustering pattern of areal subgroups exists. Non-proximal approaches to DIF detection utilizing pairwise comparisons based upon cardinal direction were unable to accurately detect either uniform or non-uniform DIF when present in this spatial structure, regardless of the DIF magnitude. The proposed method, however, was able to detect the presence of regional DIF. Two spatial clusters exhibiting DIF were simulated in locations of the study region corresponding to metropolitan statistical areas. The proposed method was able to correctly identify one of these clusters even with smaller DIF magnitudes, owing to its central location, and was able to correctly identify both clusters at higher DIF magnitudes. This observation demonstrates that the local IRT model appears most powerful for DIF detection if the DIF magnitude is of a moderate or large size and if the cluster exhibiting regional DIF is more centrally located, rather than being located on the borders or edges of the study region.

While the second case study did offer evidence that the proposed method can identify regional disparities and latent spatial trends in item functionality that are otherwise unobservable on a more aggregate level, the case study also revealed that areal units with smaller local sample sizes
were more prone to misidentification and that non-uniform DIF detection utilizing the 2PL local IRT model was somewhat limited even when spatial clustering patterns exist.

The third case study demonstrated the benefit of the proposed method when DIF is present among the areal unit subgroups but a strong spatial clustering pattern does not exist. The proposed method struggled to correctly identify areal units that exhibited DIF in this setting. Consequently, it appeared that a local approach to IRT modeling for regional DIF detection is not effective when no spatial clustering pattern of areal units exists. However, there was still an advantage to the use of the method as the visualization made possible by local modeling allowed for visual detection of regional differences that would suggest further investigation and subsequent DIF analyses.

All three simulated case studies demonstrated the overall behavior of the local IRT model. Notably, as with other local spatial models, the bandwidth is observed to have a smoothing effect with larger bandwidths corresponding to greater smoothing of the parameter estimate surface. Moreover, when employing a fixed bandwidth, smaller local calibration sizes result in higher local standard errors that gradually decrease as the bandwidth increases and local standard errors remain highest at the borders and edges of a study region.

The results from the simulated case studies demonstrated that a localized IRT modeling approach to regional DIF detection was comparable or superior to non-proximal DIF detection methods and was most powerful for detecting uniform DIF when spatial clustering and moderate to large DIF magnitudes were present. The simulated case studies suggested that the proposed method does, in fact, add to the current literature and provides a reasonable approach to investigating regional DIF that minimizes the number of pairwise comparisons necessary, incorporates spatial structure into the analysis, lessens the impact of modifiable areal unit
problems, and, with easily mappable results, provides an exploratory visual tool for such regional DIF investigations.

**Summary of Empirical Case Study Results**

While theoretical advantages and disadvantages of the proposed method were explored using the three simulated case studies, the empirical case study of a 13 item instrument designed to measure family planning method knowledge gathered from the most recent Malawi DHS survey data was provided to illustrate the local IRT model for regional DIF detection in a realistic context. Certain issues arose in practice such as non-regularity of regression points, differing local calibration sample sizes, geographic displacement that forced aggregation of data to the district level, potential violations of IRT model assumptions, and the possibility that ability distributions for the various districts differed. Some of these issues were explored in the current analysis while the possibility of others suggested future research.

For computational efficiency and owing to its potential effectiveness at detecting uniform regional DIF at the item-level, the 1PL local modeling approach was applied to the selected 13 item DHS instrument data with local model calibrations allocated to the 28 district centroids, utilizing four fixed bandwidths and a box-car kernel weighting scheme. Despite the encountered non-regularity of the regression points, the selection of several fixed bandwidths permitted an examination of the spatial relationships and of the estimated parameter surface at several levels of smoothness. While this could not entirely solve the potential problem of undersmoothing given the use of a fixed bandwidth with irregularly spaced calibration points, the use of several fixed bandwidths and various refittings of the local IRT model allowed for a visual investigation of the sensitivity of the results across different levels of spatial aggregation and disaggregation. The smoothing effect of the bandwidth seen in previous simulated case studies was still observed
and conclusions regarding general spatial trends in item functionality for two specified survey items could be made. Despite differing local calibration sample sizes, there was limited concern in this empirical example since all district samples were of adequate size for 1PL IRT modeling. Access to adequate local sample sizes might be typical if analyses are being conducted on large scale assessments such as the DHS and if data are only disaggregated to second administrative level boundaries, as this prevents local sample sizes from becoming too small. While it was not ideal to disregard individual latitude/longitude coordinates and force data aggregation to district centroids, the necessity to do so in the current example was based upon the displacement of household geographic locations for confidentiality issues. This served as a confirmation of the reasoning for regression points being commonly allocated to areal unit centroids. These realistic issues appeared to be manageable in the current context though the impact they have on modeling results and regional DIF detection should be explored further.

The use of an empirical dataset also implicitly meant that IRT model assumptions needed to be checked. The impact on regional DIF detection when using the localized approach to IRT modeling if IRT model assumption are violated remains to be properly investigated. Furthermore, the possibility that ability distributions might differ for subgroups across a spatial region of interest will also need investigated in future research to determine how differences in subgroup ability distribution impact the proposed method for regional DIF detection.

The empirical case study demonstrated the benefit of the proposed method despite these potential issues. Firstly, the application of the proposed method offered insight into spatial patterns of and regional differences in family planning method knowledge throughout Malawi, which could help guide policy makers and steer educational interventions appropriately. While knowledge of at least one family planning method in Malawi is nearly universal, research
suggests that the most commonly known methods in the country are the least effective at preventing pregnancy due to both practical and socio-cultural reasons (Dasgupta et al., 2015; John et al., 2015). Focusing on methods that might better prevent unintended pregnancies and, consequently, improve certain health outcomes for both women and children in Malawi, the use of the proposed local IRT model for regional DIF detection revealed information regarding geographic differences in IUD and implant knowledge that can assist healthcare professionals in Malawi. Although this survey instrument was most discriminating for women at lower levels of family planning method knowledge, it could be argued that this is especially where policy makers would like to affect change and where regional differences in knowledge of more effective family planning methods are most imperative. While the case study did reveal some interesting regional differences in knowledge of specific methods when matched at the same level of overall family planning method knowledge by making comparisons of local IRT models across several fixed bandwidths, it must be noted that the use of anchor items in local IRT model calibrations is recommended in order to achieve this matching when ability distributions cannot be assumed to be approximately equal across spatial subgroups or when local IRT models are to be calibrated only at small bandwidths.

There were additional benefits of using the localized approach to IRT modeling observed in this empirical case study. On the one hand, the approach might facilitate inference by researchers more familiar with Malawian family planning and contraceptive use, inevitably suggesting subsequent and further analyses. Moreover, the application of the proposed method added value to any analysis of regional differences in family planning method knowledge across Malawi due to the visualizations of both mapped IRT difficulty estimates and potential spatial subgroups suggested by significant local Moran’s I values. Lastly, the proposed method served
as a microscope wherein the local spatial patterns that contributed to significant DIF findings with non-proximal approaches could be uncovered and explored.

The results from the empirical case study complemented and supported the results from the simulated case studies, which demonstrated that a localized IRT modeling approach to regional DIF detection does provide a statistical tool to detect DIF across space without a priori groupings, identifying regional disparities and latent spatial trends that may otherwise be unobservable.

**Implications and Suggestions**

ILSAs now have a critical function influencing several policy decisions across a variety of areas. These assessments are designed with the intention to make cross-national and subnational comparisons yet items in this international context commonly exhibit DIF. While this has spurred the development of methods for detecting item-level DIF for multiple groups and while these newly developed DIF detection techniques are continually expanding to accommodate additional complexities, there is still no method or approach that currently accounts for the spatial structure existent among the groups being compared. However, the findings presented by this research support the idea that a local approach to IRT modeling for regional DIF detection could fill this void. To be clear, the findings suggest that the proposed method has the ability to uncover hidden spatial patterns and trends, to provide visualization of such spatial patterns and trends, to facilitate inference and complement non-proximal DIF detection methods, and to suggest subsequent analyses and further investigations.

Investigation of potential DIF based upon geographic or spatial location is of serious concern. It is seen as a threat to validity, limits score comparability between groups, and can also lead to inappropriate decisions that have extremely adverse effects in high stakes contexts. The
probable existence of regional DIF motivated the construction of a statistical tool that could identify regional differences in item functionality occurring across space. The proposed method has demonstrated its ability to accurately detect regional DIF in the context of ILSAs or other large-scale assessments where national and subnational comparisons are intended, especially if a strong spatial clustering pattern exists or a larger DIF magnitude is present. Additionally, the visualization offered by the proposed method appears to have great potential for policy-oriented research, driving further study and motivating reform. This localized approach to IRT for regional DIF detection could have a major impact on research across a variety of fields that utilize and rely upon the results of large scale survey instruments and future research on the proposed method should be conducted.

Limitations and Future Research

A gap in the current literature on DIF detection existed, especially in the context of ILSAs where many groups are compared across space. There was a need to propose a statistical tool for the investigation of potential regional DIF based upon geographic location without the need for pre-specified groupings and before any confirmatory DIF testing. Findings of this current work suggest that the proposed local IRT modeling approach for regional DIF detection appears to be an effective tool for identifying DIF across a spatial region, particularly if uniform regional DIF is present at a moderate to large magnitude and areal units exhibiting DIF display a spatial clustering pattern. Additionally, the proposed method offers several benefits that current methodologies for regional DIF detection cannot, notably consideration of spatial structure and visualization potential. While current findings suggest the proposed method has much promise, there are limitations to the work and several future research directions that must be explored.
The purpose of the current study was to propose, to describe, and to illustrate a novel approach for the detection of regional DIF across a spatial region without the need for a priori groupings that would also take into consideration spatial structure. Owing to the nature of the study and the newness of the proposed method for purposes of DIF detection, case studies only provided detailed examples of possible applications for the method with discussion of potential advantages and disadvantages when certain factors were manipulated. The work did not, however, include simulation studies that could provide information about overall type 1 error rates or power of the method in various contexts. As this preliminary work demonstrates that the method is capable of detecting regional DIF, future work necessarily must look at overall error rates and power. Additionally, since the method appears able to detect differences in item functionality in an empirical application, it will be of future research interest to analyze very popular ILSA datasets such as the PISA datasets in order to make comparisons between the proposed method and any previous DIF studies using these same datasets.

The application of the proposed method to the Malawi DHS data brought to light several limitations of the current approach that need to be investigated further. For instance, the non-regularity of regression points or the presence of very small local sample sizes can result in large standard errors or even in a failure of parameter estimation at specific locations when fixed bandwidths are utilized in local model calibration (Fotheringham et al., 2002, p. 46). In these circumstances, it might be better to employ an adaptive bandwidth, which would fix all local sample sizes to be equal and of an adequate size. While Fotheringham et al. (2002) argue for the use of several bandwidths for model calibration, the use of an adaptive bandwidth, especially when regression points are very irregularly spaced or when some local sample sizes are observed to be very small, is of interest and the impact of an adaptive bandwidth on the ability of the
proposed method to detect regional DIF should be explored. In addition, as in other local modeling techniques, it is of interest to offer some method for optimal bandwidth selection. Issues related to bandwidth selection should be further explored after additional study of the local IRT model itself.

The application of the local IRT model for regional DIF detection in the 13 item DHS instrument also highlighted the need for IRT model assumptions to hold. Violations of IRT model assumptions in the local IRT modeling framework have a currently unknown impact on the ability of the proposed method to detect regional DIF. Various studies examining the impact of IRT model assumptions on regional DIF detection are of interest for the future.

Relatedly, mean differences in ability across subgroups have been repeatedly shown to inflate type I error in DIF techniques (Li, Brooks, & Johanson, 2012). According to Clauser & Mazor (1998), these group ability differences could represent impact, which occurs when examinees across groups respond to an item differently because of differences on a valid construct that the item was designed to measure (i.e., the primary or target dimension). Group ability differences may cause incorrect indications of item-level DIF, increasing the type I error of a DIF detection procedure. Therefore, manipulating the group mean ability differences $\mu_d$ present would allow for an assessment of the proposed method in the presence of mean differences in ability across the regional subgroups. Since all simulated case studies have assumed equal ability distributions such that $\theta \sim N(0, 1)$, it is also of interest to explore the extent to which the proposed method suffers from increased type I errors in the presence of true mean differences in ability across the regional subgroups. However, it should be noted that the current work suggests two processes for addressing potential differences in subgroup ability distributions: (1) the comparison of DIF identified items when local IRT models are calibrated.
across varying bandwidths and (2) the use of anchor items during local IRT model calibration. Comparison of DIF identified items when estimates are calibrated across varying bandwidths is one process to address potential ability differences and was the process utilized in the current work. While not being employed in the case studies, the use of an anchor set of items during local IRT model calibration is another process that is recommended to address any potential differences in subgroup ability distributions. Regardless of the process utilized when equal ability distributions cannot be assumed, the topic of ability differences in local IRT modeling is also of great interest for the future.

Despite the limitations of the current work that suggest several future areas of research, the purpose of the current study was to describe and to illustrate a localized IRT model for regional DIF detection that required no a priori groupings and that could identify regional disparities and latent spatial trends in item functionality which may be unobservable at a more aggregate level. Findings suggest that the proposed method for regional DIF detection, utilizing a moving window approach to IRT, does offer both practical value and visual appeal when initial attempts to consider measurement invariance are being made across national, state, or other political and geographical boundaries, especially when comparisons are made to traditional DIF techniques. Consequently, the proposed local IRT modeling procedure is anticipated to provide a visual and statistical approach to DIF analyses, allowing for investigations of differences in item functionality across spatial regions that can be found in a variety of fields (e.g., education, health, economics, politics).
References


generation of international large-scale assessments. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.) *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 149-160). New York, NY: Springer.


functioning detection under the graded response model with the Raju area measures and the Lord statistic. *Educational and Psychological Measurement, 62,* 32-44.


Klieme, E. (2013). The role of large-scale assessments in research on educational effectiveness and school development. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.) *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 115-147). New York, NY: Springer.


