INVARIANT MEASUREMENT AND THE ASSESSMENT OF HOUSEHOLD FOOD INSECURITY

by

VICTORIA TAMAR TANAKA

(Under the Direction of George Engelhard, Jr.)

ABSTRACT

Food insecurity is a condition characterized by the lack of access to the food necessary for a healthy and active lifestyle. The U.S. Department of Agriculture (USDA) measures food insecurity in the United States annually with the Household Food Security Survey Module (HFSSM), an 18-item scale that references food hardships among adults and children in a household. The HFSSM is administered as a supplement to the Current Population Survey (CPS; Coleman-Jensen, Rabbitt, Gregory, & Singh, 2019). In this research, measurement is viewed by the process by which households and items are located on a line representing the latent construct of interest, household food insecurity. The five requirements for invariant measurement are: (1) item-invariant measurement of persons; (2) non-crossing person response functions; (3) person-invariant calibration of HFSSM items; (4) non-crossing item response functions; and (5) unidimensionality of the HFSSM (Engelhard, 2013). The Rasch model is an ideal-type item response theory (IRT) model that meets these requirements. It expresses the probability of endorsing an item as the function of a household's latent food insecurity and the difficulty of the item (Bond & Fox, 2015). The Rasch model was also used to calibrate the HFSSM, and the scale has been maintained over time with this model (Coleman-Jensen et al., 2019; Engelhard, Engelhard, & Rabbitt, 2017). The purpose of this dissertation is to the use properties of invariant measurement to evaluate the psychometric properties of the HFSSM. The research consists of three studies that focus on household measurement, item calibration, and dimensionality.

INDEX WORDS: Rasch measurement theory, item response theory, psychometrics, person fit, functional data analysis, model-data fit, differential item functioning, dimensionality, bifactor models, food insecurity

INVARIANT MEASUREMENT AND THE ASSESSMENT OF HOUSEHOLD FOOD INSECURITY

by

VICTORIA TAMAR TANAKA

BA, The University of South Florida, 2014

MEd, The University of Georgia, 2016

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSPHY

ATHENS, GEORGIA

2020

INVARIANT MEASUREMENT AND THE ASSESSMENT OF HOUSEHOLD FOOD INSECURITY

by

VICTORIA TAMAR TANAKA

Major Professor: George Engelhard, Jr.
Committee: Allan S. Cohen
Amanda E. Ferster
Matthew P. Rabbitt

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
May 2020

DEDICATION

This dissertation is dedicated to my mother, Vicki Denise Tanaka. Without her unending support and encouragement, I would never have been able to pursue my passions. Thank you for being the best role model and friend I could ever want.

ACKNOWLEDGEMENTS

I will always remember the first night of class in my first semester of graduate studies at the University of Georgia. Professor George Engelhard, Jr. welcomed us all to the department of Educational Psychology and shared his passion for educational measurement. “Who knows,” he joked, “maybe by the end of the semester, I’ll convert some of you to quantitative methodology.” Two semesters later, I was admitted to the doctoral program and Professor Engelhard was my advisor. I am and always will be thankful for his mentorship, guidance, and friendship. His genuine love for the field is inspirational and I am a better scholar for his influence.

I am thankful for the good fortune I have had in working with Matt Rabbitt and Amanda Ferster, who have both helped me to push myself to be a better methodologist, writer, and instructor, and Allan Cohen, who provided much-appreciated suggestions and insight that helped me in completing my work.

I would also like to acknowledge the support I felt throughout this process from the faculty and other students in Quantitative Methods. I will always be grateful for my friends and the sense of community that I have enjoyed in our program.

I want to thank AJ Thind for his advice and his enthusiasm for my work. Finally, the biggest thanks to Foye Lawrence Burch and Vivek Singh Thind for patiently listening to me think through my writing. Their support and encouragement have been invaluable as I have worked on this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This chapter introduces key concepts related to this research. I first discuss the historical development of food insecurity and the Household Food Security Survey Module (HFSSM). Then I describe the ways in which invariant measurement is defined and understood in different measurement paradigms. Finally, I review the purpose of this study, and the structure of the dissertation.

1.1 DEFINING FOOD INSECURITY

Food insecurity is a condition that exists when there is limited or uncertain access to food that is both nutritionally adequate and safe (Wunderlich & Norwood, 2006). In the United States, food insecurity is measured as a household-level construct that refers to the household's "uncertain, insufficient, or unacceptable availability, access, or utilization of food" (Wunderlich & Norwood, 2006, p. 4). This definition of food insecurity developed from work completed in the late 1980s and early 1990s and reflected the desire to estimate the number of individuals in a population who lack proper access to food. It also reflects an early decision to make separate and distinct the concepts of *food insecurity* and *hunger*. While food insecurity refers specifically to access to food, hunger is "the uneasy or painful sensation caused by a lack of food, the recurrent and involuntary lack of access to food," and "is a potential, although not necessary, consequence of food insecurity" (Wunderlich & Norwood, 2006, p. 4). By necessity, household-level food insecurity does not include all elements of individual-level food insecurity; rather, only the aspects of food insecurity that can be appropriately captured in a household-level survey are addressed in the operational definition used by the United States Department of Agriculture (USDA).

The USDA introduced the Household Food Security Survey Module (HFSSM), a part of the Food Security Supplement to the Current Population Survey (CPS), in April 1995 to estimate food insecurity prevalence on a national scale. The HFSSM includes 18 items that reference food hardships among adults and children in a household, such as anxiety regarding the amount of food in the household; experiences surrounding the lack of food, or running out; adjustments of food intake and the resulting consequences, for adults and children; and perceptions of the household's food budget and the money spent on food (Bickel, Nord, Price, Hamilton, & Cook, 2000). Items 1 through 10 are household- and adult-referenced, and Items 11 through 18 are child-referenced. Households that respond affirmatively to two items or fewer are classified as *food secure*. Households that affirm three items or more are *food insecure*. Further, households that affirm six or more items are classified as having *very low food insecurity*, and households with children have *very low food security among children* if they affirm five or more of the child-referenced items. Prior to 2006, households were also classified as *food insecure without hunger* and *food insecure with hunger* (Coleman-Jensen, Rabbitt, Gregory, & Singh, 2019). The changes to these classifications were made to reflect the fact that the HFSSM primarily measures food insecurity; the items that reference experiences of being hungry are still indicators of food insecurity because they ask about hunger due to lack of resources or access to food (Wunderlich & Norwood, 2006). The HFSSM was first calibrated using the Rasch model (Rasch, 1960/1980), and has been maintained over time using this model (Coleman-Jensen et al., 2019; Engelhard, Engelhard, & Rabbit, 2017).

In the most recent report released by the USDA, an estimated 88.9 percent of households were food secure in the United States in 2018. The remaining 11.1 percent were food insecure, with 4.3 percent of those households experiencing very low food security (Coleman-Jensen et al., 2019). Among households with children, 86.1 percent were food secure. The remaining 13.9 percent were food insecure, with 0.6 percent of those households also experiencing very low food security among children (Coleman-Jensen et al., 2019). This reflects the nature of the household-level measure: Individuals within the same household experience food insecurity to different degrees. Children may experience little to no food insecurity, while adults may face severe consequences (Coleman-Jensen et al., 2019).

1.2 INVARIANT MEASUREMENT

Invariant measurement is a perennial measurement issue that is addressed differently in the test score and scaling traditions. The test score tradition includes classical test theory (CTT), generalizability, and factor analysis. CTT depends on the observed test score, an examinee's true score, and an error score. It is assumed that the true and error scores are uncorrelated, that the average error in the population is zero, and that error on parallel tests are uncorrelated (Hambleton & Jones, 1993). The advantage to CTT is that this family of models are based on weak assumptions and have been successfully used for years, although CTT is both sample and item dependent. From the perspective of the test score tradition, *measurement invariance* occurs when individuals of equal ability from different populations receive the same observed score (Schmitt &

Kuljanin, 2008). There are different levels of measurement invariance with each successive level describing a more restrictive type of invariance:

1. Configural invariance, or identical factor structure;
2. Metric (weak) invariance, or identical factor loadings;
3. Scalar (strong) invariance, or identical item intercepts;
4. Strict factorial invariance, or identical item residual variances (Schmitt & Kulijanin, 2008).

The scaling tradition includes psychophysics and item response theory (IRT). The IRT family of models assume a single latent ability underlies a test, that the models may be applied to dichotomous data, and that the relationship between item response and person ability is determined by a one-, two-, or three-parameter logistic function (Hambleton & Jones, 1993). These parameters are item parameters and include item difficulty, item discrimination, and item guessing. IRT has stronger assumptions that are more difficult to meet. From the perspective of the scaling tradition and IRT, invariant measurement occurs when there is sample-independent measurement. In this research, measurement is viewed as the process by which households and items are located on a line representing the latent construct of interest, food insecurity. The requirements of invariant measurement are evaluated with the Rasch model, a one-parameter logistic item response model. The five requirements for invariant measurement are:

Person measurement.

1. The measurement of persons must be independent of the particular items that happen to be used for the measuring: *Item-invariant measurement of persons.*

2. A person with more ability must always have a better chance of answering an item correctly than a person with less ability: *non-crossing person response functions.*

Item calibration.

3. The calibration of the items must be independent of the particular persons used for calibration: *Person-invariant calibration of test items.*

4. Any person must have a better chance of answering an easy item correctly than a more difficult item: *non-crossing item response functions.*

Unidimensionality.

5. Items and person must be simultaneously located on a single underlying latent variable: *Wright map* (Engelhard, 2013).

Rasch explained the concept of invariant measurement as a fundamental property that *must* hold for the model to fit properly and to make comparisons between individuals and stimuli:

> "The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared. Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion" (Rasch, 1961, pp. 331-332).

That is, in order for a measure to be considered invariant, the calibration of items should not depend on the sample of persons who are administered the items. With a different sample of persons, similar item calibrations should be obtained. Because there is a duality to the IRT approach that allows for study of both items and persons, invariant measurement also requires the measurement of persons not depend upon the sample of items that have been administered. When a different sample of items are administered to the same group of people, the comparable person measures should be obtained.

1.3 PURPOSE OF THE STUDY

The purpose of this study is to use the properties of invariant measurement as outlined in Engelhard (2013) and defined by the Rasch model to evaluate the USDA's measure of household food insecurity: Household Food Security Survey Module (HFSSM). My research consists of three studies that focus on household measurement, item calibration, and dimensionality. The first study explores the use of functional data analysis and person response functions to graphically diagnose household fit. In the second study, the items of the HFSSM are examined for potential DIF and DIF-related bias with regards to the Supplemental Nutrition Assistance Program (SNAP). The third study uses bifactor models to evaluate the dimensionality of the HFSSM and discusses the potential benefits to reporting adult and child subscale scores.

The literature review in Chapter 2 is organized around the three facets of invariant measurement in the context of household food insecurity: household measurement, item calibration, and dimensionality. Terms and concepts related to the measurement of household food insecurity are also defined and discussed.

CHAPTER 2[1]

LITERATURE REVIEW

In this chapter, I present a literature review organized around the requirements of invariant measurement. First, I describe the development the U.S. household food security measure. Next, I review issues related to the requirements of invariant measurement. I start with household measurement, household fit, and applications of functional data analysis for drawing person response functions and functional clustering. Then I discuss item fit and investigating bias using differential item functioning analyses. Finally, I look at dimensionality and testing the assumptions of unidimensionality with bifactor modelling.

2.1 HOUSEHOLD FOOD INSECURITY

THE HOUSEHOLD FOOD SECURITY SURVEY MODULE

As mentioned in the previous chapter, the measurement of food insecurity developed from work completed in the 1980s and early 1990s. The guiding principle for the research that would culminate in the creation of a national food-insecurity scale was stated in the President's Task Force on Food Assistance (1984):

> "It has long been an article of faith among the American people that no one in a land so blessed with plenty should go hungry. Both in our private associations and in our public policy we have incorporated the idea that we as a community have a

[1] Portions of this chapter were published in the *Journal of Applied Measurement* (Tanaka, Engelhard, & Rabbitt, 2019) and were presented at the annual meeting of the National Council on Measurement in Education in New York, NY (Tanaka, Engelhard, Rabbitt, & Jennings, 2018), and Toronto, Ontario (Tanaka & Engelhard, 2019). Reprinted here with permission of the publisher.

> moral and human obligation to assist those in distress. … Hunger is simply not acceptable in our society" (p. 2).

The purpose of the report was to analyze the state of food assistance programs and to understand hunger in the United States. It was noted that the terms *hunger*, *poverty,* and *unemployment* are used interchangeably, and hunger was defined as both "a condition in which the level of nutrition necessary for good health is not being met because one lacks access to food" (Task Force, 1984, p. 3) and "a situation in which someone cannot obtain an adequate amount of food, even if the shortage is not prolonged enough to cause health problems" (Task Force, 1984, p. 3). The Task Force also found that there was no "official 'hunger count' to estimate the number of hungry people, and so there are no hard data available to estimate the extent of hunger directly" (Task Force, 1984, p. 37), though clearly "there is hunger in America, […] an intolerable situation" (p. xv). This became the impetus for research that led to the Household Food Security Survey Module (HFSSM).

The food insecurity project began in 1992 when, under the direction of the Ten-Year Comprehensive Plan for the National Nutrition Monitoring and Related Research Program, established by an act of Congress in 1990, researchers were tasked with developing a standard measure of food insecurity that could be used at the national and at state levels (Bickel et al., 2000). The result was the HFFSM, first administered in April 1995 as a supplement to the Current Population Survey (CPS). Following this first administration, a numerical scale and food security categories were developed to describe the food security status of households in the United States within the preceding 12-month period. Following this success, the HFSSM has been administered annually each year since 1995 and has proven to be both stable and valid over time (Bickel et al., 2000). A summary of food insecurity prevalence estimates based on responses to the HFSSM is published, so that it is possible now both to track changes in food insecurity over time and to estimate the number of households experiencing hunger in the United States—addressing the problem first noted in the President's Task Force on Food Assistance.

The HFSSM is an 18-item scale that references food hardships for the household, and for adults and children within the household. The full scale is presented in Table 1. These items are coded dichotomously for analysis; responses in bold indicate an affirmative response. Households are assigned a food security status classification based on their raw score to the scale. Households with a score of two or fewer are food secure, while households with a score of three or higher are food insecure or food insecure among children with a score of two or more of the child-referenced items. Households are further classified as having low food security or very low food security if they receive a raw score of six or higher in households without children or eight or higher in households with children. The difference in low and very low food security is in the characteristics of the households: Households that have low food security have poorer quality of diet but report fewer instances of a reduction in food intake. In contrast, households that have very low food security have both poorer quality of diet as well as disrupted or reduced food intake (Coleman-Jensen et al., 2019). Finally, households have very low food security among children if they respond affirmatively to five or more of the child-referenced items. For a full summary of the HFSSM, including its development and administration, see Bickel et al. (2000).

In 2006, the Committee on National Statistics reviewed the measure for potential improvements. Their recommendations are presented in Table 2 (Wunderlich & Norwood, 2006, pp. 9-11). The USDA reorganized the HFSSM shortly after this report was released. Whereas prior to 2006, the items were organized by item difficulty, after 2006, items were grouped by into adult- and child-referenced groups, according to findings from related cognitive research (Wunderlich & Norwood, 2006). Nord (2012) also conducted a study to address Recommendations 5-1 and 5-2, concluding that the Rasch model is appropriate for the data collected with the HFSSM and that there is no differential item functioning (DIF) for households with and without children. Similar results have been found with other subgroups (Rabbitt, 2018; Rabbitt & Coleman-Jensen, 2017). Following Recommendation 5-4, alternate methods of food security classification have also been examined, with a small improvement in household classification with the experimental classification systems (Coleman-Jensen, Rabbitt, & Gregory, 2018; Nord, & Coleman-Jensen, 2014).

THE SUPPLEMENTAL NUTRITION ASSISTANCE PROGRAM

Food assistance programs exist to increase the availability of nutritious foods to eligible low-income households, and to reduce their food insecurity. SNAP is the largest food assistance program in the U.S. To be eligible to participate in SNAP, households must meet financial, work-related, and categorical conditions (Fox, Hamilton, & Lin, 2004). Paradoxically, food insecurity has been found to be *more* prevalent in households receiving SNAP benefits than in low-income households without assistance (Gundersen & Oliveira, 2001; Nord & Golla, 2009). Research into the effects of food assistance programs has not uncovered any mechanism through which SNAP benefits increase a household’s food insecurity. Instead, it is understood that the relationship is more complex than it seems at first glance. Households are not constantly food secure or food insecure. When a situation is bad enough for households to seek assistance, the hardships they face are sufficiently severe enough to overwhelm the beneficial effects of SNAP on food insecurity (Wilde & Nord, 2005). However, attempts to study the relationship between SNAP and food insecurity are complicated by unmeasured or unobserved characteristics that are likely correlated with both SNAP and food insecurity (Gibson-Davis & Foster, 2006).

With SNAP, households are presumed to have enough resources to purchase up to 103 percent of the Thrifty Food Plan (TFP), which is the basis of SNAP (Gundersen & Oliveira, 2001). The TFP is one plan of four that the USDA created for different cost levels—Thrifty, Low-cost, Moderate-cost, and Liberal. The TFP reflects dietary recommendations, food consumption patterns, and food prices as calculated by the USDA’s Center for Nutrition Policy and Promotion (CNPP; Carlson, Lino, Juan, Hanson, & Basiotis, 2007). Food prices that are used to calculate the TFP are the same prices that many low-income families pay for food. The TFP may be considered a national standard for the minimal cost required for a healthy, nutritious diet. The weekly and monthly cost for each plan is calculated every month. Costs are given for individuals in four-person families, broken down by age and gender into fifteen specific age-gender groups. By adjusting food costs for each person in the household and then summing these costs, the cost of food for the household as a whole may be calculated. The TFP, and the monthly cost of food report for each of the USDA food plans, provide a realistic reflection of the availability and price of food, and are tools for understanding hunger and food security in the U.S.

Household food insecurity categories are assigned based on sum scores from the household food insecurity scale. These classifications do not depend on which items have been affirmed. This means that when SNAP-based differential item functioning (DIF) is present in the items that are used to calculate the sum score, the resulting bias can vary across the food insecurity categories. Households who are classified as food insecure and who affirmed these items may experience even greater food insecurity than suggested by the classification.

Misreporting in surveys can also introduce measurement error. In fact, large, systematic misreporting can make it difficult to answer questions about important programs such as SNAP with survey data alone (Mittag, 2013). Analysis of misleading survey data can lead to misunderstanding of program effects on the target population. With respect to SNAP, this misreporting can be very severe and cannot be ignored (Mittag, 2016). For example, with the Current Population Survey (CPS), it has been found that as much as 26 percent of households that do receive SNAP benefits falsely report that they do not when surveyed (false negative), while the number of households that do not receive SNAP benefits but report they do (false positive) is only 1.2 percent (Mittag, 2016). This has implications for the SNAP-based DIF study. If the false negative rate is so high, then there is the potential that inaccurate survey reports are the cause of the detected DIF, and that the bias noted between SNAP and non-SNAP groups is a result of misreporting. Mittag (2016) suggests several methods for addressing misreporting of benefits including:

- Calculating the reporting rating and scaling survey estimates up accordingly,
- Relying on determinants of SNAP eligibility rather than self-reported participation, and
- Using statistical methods to correct for measurement error.

2.2 HOUSEHOLD MEAUSREMENT

Person response functions (PRFs) represent the functional relationship between the probability of a person endorsing or answering correctly a set of items. The person and the items are located on a continuum with estimated locations based on a measurement model. For example, the PRFs for the Rasch model (Rasch, 1960/1980) are assumed not to cross, one of the major requirements of invariant measurement (Engelhard, 2013). When this assumption is violated and PRFs do cross, then ordering of people vary around the intersection points, creating differential ordering of persons and leading to difficulty in substantive interpretations of person performance (Perkins & Engelhard, 2009). Other IRT models with additional item parameters can also be used to define PRFs (Baker & Kim, 2004). PRFs are closely related to person fit with numerical indices of person fit requiring an examination of PRFs to diagnose the sources of misfit. It is well known that misfitting items can be identified with various statistical indices, but a visual examination of the item response functions is still useful (Wells & Hambleton, 2016). Similar recommendations apply for evaluating person misfit.

PRFs are in many ways the mirror image of item response functions with a focus on each person rather than an item. This duality between persons and items has a long history with key insights provided by Mosier (1940, 1941). Carroll (1985) used person characteristic functions as a means of defining ability in terms of item difficulty and the individual's probability of success, therefore allowing for valid inferences about the ability being measured. Trabin and Weiss (1979) described the use of trace lines for modeling an examinee's response to test items. The slope of the trace line or person response function represents the consistency of the responses (analogous to item discrimination), and the inflection point represents the item difficulty for an examinee of a given ability level. Lumsden (1977) later built on this work by using what he called person characteristic curves to compare individual differences, such as person reliabilities and aptitudes.

The research described above is grounded in PRFs based on parametric models, and therefore must satisfy certain requirements (e.g., PRFs must be monotonically increasing). PRFs based on non-parametric models have several advantages, such as not requiring increasing monotonicity—leading to the potential identification of different types of person misfit, such as guessing or slipping (Walker, Jennings, & Engelhard, 2017). Much recent work with PRFs, then, has seen more exploration of these non-parametric methods for modeling empirical person response functions, such as the use of FDA and functional clustering (Turner, 2018) and the Tukey-Hann approach to smoothing PRFs (Jennings & Engelhard, 2017).

FUNCTIONAL DATA ANALYSIS

Functional data analysis (FDA) has been used in several psychometric contexts (Ramsay, 1997, 2016; Ramsay & Silverman, 2005), but it has not been used to represent PRFs as proposed in this study. Essentially, FDA is a philosophical and statistical approach that treats functions as data points. FDA can be used to represent discrete data, such as item responses, as a series of smooth and continuous functions. FDA builds on the principles of multiple linear regression, using linear combinations of mathematically independent basis functions to represent functions. The FDA equation for the linear combination of K functions is generally expressed as

$$x(t) = \Sigma_{k=1}^{K} c_k \phi_k(t)$$

where c_k represents the coefficients and ϕ_k represents the basis functions. Spline basis functions are commonly chosen for non-periodic functional data without strongly cyclical variation (Ramsay & Silverman, 2005). These functions are piecewise polynomials that are defined by their order, and by the location of the knots that join the polynomials together. B-splines developed by de Boor (2001) are the most popularly used in the literature (Ramsay & Silverman, 2005).

In order to detect patterns in the person response functions obtained from FDA, cluster analysis can be used to group data that have some commonalities. Clustering is particularly useful in aiding in interpretation of results for PRFs for large datasets. Many methods exist for clustering data. In the case of FDA, the primary goal is to cluster functional data according to some criteria. Tarpey and Kinateder (2003), for example, have discussed the process of finding curves that are representative of homogenous subgroups based on modes of variation. Other methods have also been proposed for the functional clustering of sparse data (James & Sugar, 2003), time series data (Fröhwirth-Schnatter & Kufmann, 2008), polynomial regression models (Samé, Chamroukhi, Govaert, & Aknin, 2011), and high dimensional data clustering (Bouveyron & Jacques, 2011). The approach developed by Bouveyron, Côme, and Jacques (2015) is used in this study as a model-based functional clustering method that visually represents clustered curves. The R package FunFEM clusters curves drawn by FDA into *k* homogenous groups (Bouveyron, Côme & Jacques, 2015).

2.3 ITEM CALIBRATION

As previously mentioned, the Rasch model has strict requirements for invariant measurement (Engelhard, 2013). After unidimensionality has been checked, model fit statistics are used to determine the extent to which the requirements of invariant measurement are met. The most common measures used to diagnosis fit are Infit Mean Square and Outfit Mean Square. Infit Mean Square is a weighted fit statistic of response patterns that is less sensitive to extreme responses. Outfit is an unweighted fit statistic that is sensitive to outliers, or person responses that are distant to item difficulty. Both Infit Mean Square and Outfit Mean Square have an expected value of 1.0. Small Infit values could indicate a Guttman response pattern, while large values could indicate an aberrant pattern (Linacre, 2015). Large Outfit values may indicate outliers, while small values may indicate overly predictable responses (Linacre, 2015). Traditionally, fit statistics between 0.8 and 1.2 have been considered good, although Smith, Schumacker, and Bush (1998) have discussed the history of item fit statistics and suggested the use of various standardized fit statistics. As pointed out by Engelhard and Wind (2018), it is also possible to use general fit categories (A, B, C, and D) for interpretative purposes. The recommended guidelines are presented in Table 3.

DIFFERENTIAL ITEM FUNCTIONING

The Rasch model represents an ideal of how a scale should function. Ideals are always violated in practice. However, most violations are inconsequential, even when they are statistically significant (Linacre, 2012). Differential item functioning (DIF) analyses can be useful in determining when these violations require further attention. When the conditions of the Rasch model are met, it is possible to assess the differences between groups using a DIF analysis. In general, such an analysis identifies a focal group and a reference group. The focal group is the group of interest, and the group which could possibly be disadvantaged by the item. The reference group is used for the purposes of comparison (de Ayala, 2009). A DIF analysis provides insight into how an item is experienced by the reference and focal groups. When no DIF, or non-significant DIF, is detected then group membership does not matter because the items function in a comparable way for everyone. If there is a statistically significant difference, members of the disadvantaged group may feel substantive consequences, depending on the purpose of the scale or test (Linacre, 2012).

2.4 DIMENSIONALITY

BIFACTOR MODELS

Bifactor measurement models provide an approach for examining and modeling construct-relevant multidimensionality (Reise, 2012) of the HFSSM heretofore used in the measurement of food insecurity. The bifactor model was proposed by Holzinger and Swineford (1937) and extended to item-level bifactor analyses by Gibbons and Hedeker (1992). In the bifactor measurement model for items, the covariances (or correlations) among items can be accounted for by a general factor underlying all the items, and several specific or residual factors. Conceptually, the general factor is the “broad target construct” that the instrument was created to measure (Reise, 2012, p. 668). In the context of this study, the broad target construct that the HFSSM was created to measure is household food insecurity. The specific factors account for clusters of items with similar content, such as the adult and child items in HFSSM. They can be considered conceptually narrower "subdomain constructs" (Reise, 2012, p. 668).

The bifactor model is a potentially useful way to represent construct-relevant multidimensionality in cases where broad constructs have multiple subdomains that may be of interest. In the case of food insecurity, measured by the HFSSM, the unidimensional Rasch model is typically used for the broad construct of household food insecurity. The two distinctive subsets of items related to household/adult and child experiences with food insecurity are examples of narrower subdomain constructs. All the domains of food insecurity are used for policy design and evaluation; however, household food insecurity is the most referenced (e.g., Coleman-Jensen et al., 2019). It is common to design instruments with subdomains or subscales when developing unidimensional scales (Wilson, 2005), and it is important to consider whether meaningful subscales that reflect multidimensionality can be constructed. In summary, bifactor models have several desirable characteristics, including an approach for examining latent variable scores on several dimensions (Reise, 2012). It should be stressed that from the perspective of bifactor measurement models the specific factors are essentially residual factors that are estimated after the general and dominant factor (DeMars, 2013).

In Chapter 3, I present a study that examines person measurement and fit in the context of household food insecurity. This is accomplished using FDA to model and functionally cluster PRFs within levels of household food security status classifications as defined by the USDA, and the fit statistics groupings as described in Engelhard and Wind (2018).

CHAPTER 3[2]

HOUSEHOLD MEASUREMENT

This chapter describes an idiographic approach to household measurement. Household item response patterns are used to explore within-household variability. Functional data analysis (FDA) techniques are used to draw and cluster person response functions (PRFs) within levels of food security status classifications and fit statistic groupings. This study illustrates the assertion that, in addition to traditional numerical summaries, graphical displays are necessary to understand and diagnose household misfit.

3.1 INTRODUCTION

An idiographic approach to scientific inquiry explores individual differences and within-person variation, while a nomothetic approach focuses on group-based differences and between-person variation. Allport (1937) characterized the nomothetic perspective as a search for general laws that hold across persons, while the idiographic perspective explores specific individual cases for which general laws may or may not be established. According to Molenaar (2004),

> Psychometrics and statistical modeling as we now know it in psychology are incomplete. What is lacking is the scientific study of the individual, his or her structure of IAV [intraindividual variability], for its own sake. Scientific

[2] An early version of this chapter was presented at the annual meeting of the National Council on Measurement in Education in Toronto, Ontario (Tanaka & Engelhard, 2019).

psychology can only become complete if it includes the idiographic point of view, alongside the nomothetic point of view (p. 216).

Molenaar stresses that an individualistic and idiographic approach to measurement should be given the same level of consideration as the traditional group-based nomothetic perspective. It is important to examine person fit as well as item fit when assessing the fit of a model and identifying misfit.

Item fit is typically evaluated in psychometric studies. It is well known that good model-data fit is necessary in order to realize the desirable invariance properties for item response theory models (Engelhard, 2013). Wells and Hambleton (2016) provide a description of how to approach item fit from the perspective of residual analyses. However, in addition to item fit, it is necessary to evaluate person fit in order to validate the response patterns for persons (Messick, 1995; Wright, 1980). Examples of person misfit may include difficulties in understanding the items, lack of motivation in responding, multidimensionality, person unreliability, response bias, and extreme responses (Ferrando, 2015).

Person response functions (PRFs) can serve as a means of bringing the person back into measurement as a part of an idiographic approach (Molenaar, 2004). Person fit indices provide numerical statistical summaries of person misfit, but graphical displays such as PRFs are also necessary to accurately diagnose sources of aberrant responses, such as within-person dimensionality. Graphical displays play an important role in examining item misfit, but the use of graphical displays related to person fit has not been stressed enough in the literature on psychometrics. PRFs offer an approach for examining person responses. Despite their promise in understanding person misfit, PRFs can be difficult to evaluate, particularly in the context of large-scale assessments when there are many more persons than items.

The purpose of this study is to introduce functional data analysis (FDA; Ramsay, 1997) as an approach for examining household fit with PRFs. Our approach includes the use of cluster analyses of the functional representations of the PRFs to categorize patterns of aberrant person responses. Response data from the Household Food Security Survey Module (HFSSM) are used to illustrate this approach.

3.2 METHODOLOGY

ANALYSES

Earlier analyses of household fit for these data were based on the Rasch model as a point of reference for interpreting misfit. The dichotomous Rasch model is a logistic item response theory (IRT) model that describes the relationship between person and item locations on a latent variable. The log odds of a person endorsing an item can be expressed as follows:

$$\ln[P_{ni1}/P_{ni0}] = \theta_n - \delta_i$$

where:

P_{ni1} is the probability of person n endorsing item i,

P_{ni0} is the probability of person n not endorsing item i,

θ_n is the ability of person n, and

δ_i is the difficulty of item i.

The Rasch model has strict requirements for invariant measurement including: (1) item invariant measurement of persons, (2) person invariant calibration of items, (3) non-crossing PRFs, (4) non-crossing item response functions, (5) and unidimensionality

(Engelhard, 2013). The Rasch analysis was conducted using the Facets software developed by Linacre (2015).

PRFs were analyzed in R using the FDA package (Ramsay, 2011). There are several decisions required for the construction of PRFs with this package. The three main decisions include setting the location of the knots, selecting the order of the splines, and determining the number of basis functions to be used in estimating the PRFs. In this study, 18 knots were selected to correspond to the item difficulties of the 18 items in the HFSSM, and an order of four was used that reflects a third-degree polynomial for the splines. Based on the number of knots and order, the number of basis functions required is 21. Finally, b-splines served as the basis functions.

An example of an expected PRF created with FDA is presented in Figure 1 (Panel A). This PRF reflects a perfect Guttman scale. A person's probability of providing a correct answer decreases as a function of increasing item difficulty (scaled in logits). Panel B in Figure 1 provides an example of an unexpected PRF created with FDA. As item difficulty increases, initially the individual’s probability of a correct response decreases. However, at higher levels of item difficulty, there is an unexpected increase in the probability of a correct answer that is indicative of person misfit. The misfit in Panel B suggests potential within person multidimensionality.

The observed PRFs were grouped in R using the funFEM package (Bouveyron, Côme, & Jacques, 2015). This package allows for functional clustering of data based on a discriminative functional mixture model. As with the FDA package, the funFEM package requires the specification of several options. First, the number of clusters is selected for grouping the PRFs. Based on the Akaike Information Criterion (AIC) and Bayesian

Information Criterion (BIC) indices, four clusters were suggested as a reasonable number of acceptable clusters to identify person misfit. Next, the discriminative latent model (DLM) for estimating the clusters is selected. Through model selection criteria, the $\Sigma_k\beta_k$ (also called D_kB_k) was found to be most appropriate for this data (Turner, 2019). Finally, the algorithm is selected—either randomly, using *k*-means, or hierarchical clustering. Hierarchical clustering was used for these data. Future research is needed regarding decisions that are made in using FDA with assessment data.

PARTICIPANTS

This study uses data on U.S. households who responded to the HFSSM in 2012 to 2014. Households were below 185 percent of the federal poverty line and had at least one child under the age of 18 (N = 7,324). These data were analyzed earlier by Engelhard, Rabbitt, and Engelhard (2017).

3.3 RESULTS

Rasch summary statistics for the HFSSM data are presented in Table 3. The Rasch analysis explained 60.91 percent of the variance. Households had an average food-security measure of -2.63 (SD = 2.28), indicating lower levels (in terms of the severity of food hardship) of food security. The reliability of household separation is .84. The average Infit was 0.99 (SD = 0.56) and the average Outfit was 0.73 (SD = 1.23). Table 4 provides a summary of the household level person fit statistics based on four categories. Approximately 22 percent of households (N = 1608) had poor Infit (categories C and D), and approximately 10.3 percent (N = 756) had poor Outfit (Categories C and D. Households fit statistics of 1.50 or greater are considered unproductive for measurement (Engelhard, Rabbitt, & Engelhard, 2017).

Using the procedure outlined in the previous section, PRFs were drawn for each household, and then clustered. To facilitate interpretation, households were first grouped by their severity of food insecurity: Food Secure, Low Food Security, and Very Low Food Security. These levels are based on the USDA's classification thresholds for households with children. Households with children that endorse two or fewer items are considered *food secure*, while households that endorse three or more items are *food insecure*. Additionally, households that endorse between four and seven items have *low food security*, and households that endorse eight or more items have *very low food security* (Coleman-Jensen et al., 2019).

The PRFs by food insecurity level are presented in Figure 2. Each figure includes all the PRFs for a level, as well as the clustered display of PRFs. Separate PRFs are difficult to interpret, and the clustered PRFs provide more information. Households with low food security appear to exhibit some degree of multidimensionality, exhibiting the PRF shape shown in Panel B (Figure 1). The two other categories (food secure and insecure) do not suggest any issues with misfit. Households are also grouped by fit categories in Figure 3. These categories are presented in Table 5. As was the case with Figure 3, it is difficult to interpret individual PRFs. Based on four clusters, Category C households (unproductive for measurement, but not distorting of measures) appear to have the most variation between categories.

3.4 DISCUSSION

This study describes an approach for examining household fit with person response functions (PRFs) that are estimated with functional data analysis (FDA). One of the major advantages of using FDA is that non-monotonic curves can be displayed that identify potential sources of within-person multidimensionality. Another advantage is that the PRFs can be viewed holistically as data points, and then categorized with various types of cluster analyses that have been developed for FDA.

It is important to stress that although item fit is routinely assessed when examining model-data fit, person fit should also be considered (Ferrando, 2015; Messick, 1995; Wright, 1980). Good model-data fit for items and persons is needed in order to realize the advantages of invariant measurement. In the illustrative example on food insecurity, 22 percent of households had poor Infit and approximately 10 percent had poor Outfit. FDA with PRFs offers a tool for diagnosing sources of misfit that go beyond numerical indices of person fit. The displays of PRFs and the clustering of PRFs offer a useful tool for visually evaluating person fit. As with other graphical displays, the researcher is still faced with a variety of decisions regarding the interpretation of these displays. Numerical summaries of person fit are inadequate for fully understanding how people misfit statistical models, but FDA is a useful tool for estimating and clustering PRFs in order to identify persons who require additional analyses to determine the sources of misfit and other characteristics of persons who are clustered together.

This study takes an ideographic approach to measurement by emphasizing PRFs as a graphical method for evaluating person fit. It also suggests the potential uses of FDA for estimating those PRFs. By drawing and examining PRFs, individual misfit can be diagnosed, and then individual profiles reflected in PRFs can be clustered to provide a basis for future research on sets of persons that may exhibit aberrant responses. In practical assessment situations, it is also important to examine why certain groups have related patterns of misfit. Future research should consider selecting different options with the FDA analyses—for example, the use of basis functions other than b-splines, or alternate classification algorithms for estimating clusters. An ideographic approach highlights the importance of the individual fit in social science measurement.

Chapter 4 examines issues regarding item calibration. Differential item functioning (DIF) analyses of the HFSSM reveal potential biases in the scale with regards to participation in food assistance programs. The implications of these results for policy are discussed.

CHAPTER 4[3]

ITEM CALIBRATION

In this chapter, I present a study that examines the Household Food Security Survey Module (HFSSM) for potential differential item functioning (DIF) with respect to participation in the Supplemental Nutrition Assistance Program (SNAP, formerly the Food Stamp Program), the largest food assistance program in the United States. Previous research has noted that food insecurity tends to be more prevalent in households receiving assistance than in households without assistance (Gregory, Rabbitt, & Ribar, 2015; Gundersen & Oliveira, 2001; Nord & Golla, 2009).

4.1 INTRODUCTION

Prior research suggests that households who participate in the Supplemental Nutrition Assistance Program (SNAP) report higher levels of food insecurity (Coleman-Jensen et al., 2019; Gundersen & Oliveira, 2001; Gregory, Rabbitt, & Ribar, 2015; Nord & Golla, 2009; Wilde & Nord, 2005). In this study, differential item functioning (DIF) in the Household Food Security Survey Module (HFSSM) related to participation in SNAP is examined. The following research questions are addressed:

1. Are there subgroup differences in reported food insecurity based on SNAP participation within the previous 12 months?
2. Is there DIF related to SNAP participation within the previous 12 months?

[3] An earlier version of this chapter was published in the *Journal of Applied Measurement* (Tanaka, Engelhard, & Rabbitt, 2019). Reprinted here with permission of the publisher.

3. Do the rates of aberrant household responses differ based on SNAP participation?

An earlier study (Rabbitt, 2018) of DIF in the HFSSM related to SNAP receipt, based on the 8 child food insecurity items, detected DIF for Item 16 (*child(ren) skipped meals*). If DIF is present, then the HFSSM could produce inaccurate food insecurity prevalence estimates when applied to households based on SNAP participation. We also examine how household model-data fit is related to SNAP participation (Engelhard, Rabbitt, & Engelhard, 2017).

4.2 METHODOLOGY

PARTICIPANTS

This study uses data on all households responding to the HFSSM in 2015 and 2016 with income below 185 percent of the federal poverty line, and who had at least one child under the age of 18 and participated in the Current Population Survey of the US census. Households with extreme responses (zero or 18) were excluded. Households with invalid information on SNAP receipt indicators were also excluded. Additionally, households in Alaska and Hawaii were excluded because of differences in SNAP administration. The sample size was 3,931 with 2,257 (57.4 percent) reporting participation in SNAP, and the remaining 1,675 households (42.6 percent) not reporting SNAP participation.

DIFFERENTIAL ITEM FUNCTIONING MODELS

The Facets computer program (Linacre, 2015) was used to calibrate and compare items based on the household's SNAP participation status. The Rasch measurement model is useful for examining the relationship between the severity of food insecurity and item difficulty. Households and items are ordered according to proficiency and difficulty along the latent variable. The log odds of a respondent endorsing an item for Model I can be expressed as follows:

$$\ln\left[\frac{P_{nij1}}{P_{nij0}}\right] = \theta_n - \delta_i - \Delta_j \qquad (1)$$

where:

P_{nij1} = probability of respondent n endorsing item i from SNAP subgroup j,

P_{nij0} = probability of respondent n not endorsing item i from SNAP subgroup j,

θ_n = logit-scale location of respondent n,

δ_i = logit-scale location for item i, and

Δ_j = logit-scale location of SNAP subgroup j,

Model I is a facets model with three facets: household, item and SNAP subgroup.

The log odds probability for Model II can be expressed as follows:

$$\ln\left[\frac{P_{nij1}}{P_{nij0}}\right] = \theta_n - \delta_i - \Delta_j - \delta_i\Delta_j \qquad (2)$$

Model II is the three-facet model with an added interaction term ($\delta_i\Delta_j$) that reflects potential item by SNAP participation interaction effects. It should be noted that Model II emphasizes the interaction effects. To control for the connectedness of the subsets formed by the SNAP subgroups, the SNAP facet was included as a dummy facet anchored at zero, and this anchoring allowed for examination of the interaction effects, following recommendations by Linacre (2012).

4.3 RESULTS

The three-facet Rasch model explained 60.3 percent of the variance in the data, and this supports the inference that the scale is unidimensional (Bond & Fox, 2015). The

DIF analysis explained a very small amount of the additional variance (0.12 percent). The Wright map (Figure 4) is a visual representation of household food insecurity and item difficulty. The distribution is positively skewed, with more households falling in the lower on the logit scale (i.e., having lower levels of food insecurity). Generally, the child-referenced items appeared to be more difficult to endorse than the adult-referenced items: most of the child items appear higher up on the map (i.e., higher levels of food ins than the adult items. This suggests that these items were harder to endorse and represent greater severity in food-insecure conditions.

Rasch summary statistics for the data are presented in Table 6. Overall, Infit Mean Square and Outfit Mean Square statistics were good for the households, items, and SNAP facets. Households had an average food insecurity level of −2.78 (SD = 2.25) logits. Items had an average difficulty of 0.00 (SD = 3.06) logits. The SNAP facet was anchored at zero. The reliability of separation was high for households [R = 0.83; χ^2 (3930) = 24266.4, $p < .05$]. The reliability for households corresponds to the traditional coefficient alpha in classical test theory. The items [$R > .99$, χ^2 (17) = 27164.2, $p < .05$] also had a significant reliability of separation index.

Summary statistics for items are given in Table 7. Labels correspond to those used to identify items in the Wright map (Figure 4). Infit Mean Square and Outfit Mean Square statistics are given, as well as classifications of the items for fit statistics. The categories are based on recommendations made by Engelhard and Wind (2018) for interpreting the fit statistics calculated in a Rasch context. Generally, for all items, Infit Mean Square was close to the expected value, or fell within category A. Infit Mean Square was the lowest for Item 4 [*adult(s) cut size or skipped meals*; Infit Mean Square =

0.75] and highest for Item 11 (*relied on low-cost foods for children*; Infit Mean Square = 1.33). Outfit Mean Square was close to the expected value (within the A category) for most of the items. It was worst for Item 2 (*food bought would not last*; Outfit Mean Square = 2.51), Item 3 (*could not afford to eat balanced meals*; Outfit Mean Square = 1.59), and Item 11 (*relied on low-cost foods for children*; Outfit Mean Square = 1.65).

Participation in SNAP within the previous 12 months (a response of "Yes," coded 1) was the focal group, and nonparticipation in SNAP (a response of "No," coded 2) was the reference group for the DIF analysis (Table 8). The first three items of the HFSSM apply to the household in general. There were significant differences ($t = -6.26$, $p < .0001$; $t = -3.61$, $p = .0003$; and $t = 6.29$, $p < .0001$) between groups for these items. Food insecurity was higher for the SNAP group on Item 1 (*worried food would run out*) and Item 2 (*food bought would not last*), and higher for the non-SNAP group on Item 3 (*could not afford to eat balanced meals*). Among the adult items, there was no difference in food insecurity between groups, but there was a significant difference for two of the child items. The difference was significant for Item 11 (*relied on low-cost foods for children*; $t = 2.18$, $p = .0295$), and for Item 12 (*could not feed children balanced meals*; $t = 3.30$, $p = .0010$). For both items, food insecurity was higher for the non-SNAP group.

It is interesting to note that DIF was balanced in the scale. For example, nine of the items were higher in the SNAP group, while nine of the items were lower in the SNAP group. This resulted in the mitigation of the potential biasing effects of DIF in the use of the scale in this context. The impact of item level invariance is also minimally related to the classification of food insecurity because the overall policy implications are made based on counts of items rather than cut scores based on the latent variable scale.

Table 9 presents the description of the household fit analyses. Summaries of household fit are presented separately for the Infit and Outfit statistics, as well as by participation in SNAP. The Infit statistics are similar for respondents across SNAP subgroups with an overall misfit of 5.0 percent. As expected, the Outfit statistics indicate a slightly higher rate of overall misfit (6.9 percent) and the misfit rates are also comparable over SNAP subgroups. Household misfit was comparable in both groups. The overall rates of household misfit are similar to those found in earlier research (Engelhard, Rabbitt, & Engelhard, 2017).

4.4 DISCUSSION

The purpose of this study was to examine differential item functioning (DIF) on the Household Food Security Survey Module (HFSSM) used to evaluate food insecurity in the United States. The HFSSM is an 18-item scale, and it is used extensively to monitor policy related to food insecurity. It is important to examine measurement invariance with DIF analyses related to participation in Supplemental Nutrition Assistance Program (SNAP).

First, subgroup differences based on reported food insecurity related to SNAP participation within the previous 12 months were examined. Results confirm earlier research that indicated that households receiving SNAP report higher levels of food insecurity. In this study, the substantive reasons for the unexpected direction of this relationship were not examined; however, Gregory, Rabbitt, and Ribar (2015) provide an extensive discussion of this topic.

Next, the HFSSM was examined for DIF related to SNAP participation within the previous 12 months. The overall item fit analyses are good based on the Infit Mean

Square statistics. The Outfit Mean Square statistics identified one item (Item 2: *Food bought would not last*) as misfitting. The results of the DIF analyses indicated that 5 out of 18 items have statistically significant DIF. One of the interesting findings is that the direction of the DIF appears to minimize the overall influence of DIF—in other words, about half of the items exhibiting DIF went in one direction, while the other half went in the opposite direction. This finding reduces the potential inaccuracy in estimates of overall household food insecurity when continuous measures are used in empirical analyses of the relationship between SNAP and food insecurity. It is also important to note that levels of food insecurity are based on sum scores, and this may also minimize the biasing effects of DIF on the substantive and policy decisions. Additional empirical research is needed to explore the issue of how DIF affects the categorization of households into food security status categories because these decisions are based on sum scores.

Finally, a research question related to how rates of aberrant household responses differ based on SNAP participation was addressed. The data indicates that the rates are similar in households regardless of SNAP participation. Further research is needed to identify why approximately 5.0 percent to 6.9 percent of households have aberrant response patterns as found in earlier research (Engelhard, Rabbitt, & Engelhard, 2017).

In summary, SNAP is an important line of defense against food insecurity and poverty in the United States. The psychometric quality of the HFSSM is of critical value to the national debate regarding food insecurity in the United States. Accurately estimating food insecurity has multiple implications for the quality of life for many people. The DIF analysis explained a very small amount of the additional variance (0.12

percent) in this study after controlling for the main effects. Nevertheless, future research should still consider why specific items appeared to exhibit measurement invariance related to SNAP participation. One very promising area for future research is to focus on the rate of aberrant responses. Each household has a unique response pattern for the items included in the HFSSM, and there may be additional insights that can be gained by using person response functions to explore these patterns (Engelhard, Rabbitt, & Engelhard, 2017).

In Chapter 5, the dimensionality of the HFSSM is examined using bifactor models to determine the utility of separate subscales for the adult- and child-referenced items. These multidimensional models are compared to two unidimensional item response theory (IRT) models: a one-parameter logistic model and a two-parameter logistic model.

CHAPTER 5[4]

DIMENSIONALITY

In this chapter, I present a study that examines the dimensionality of the Household Food Security Survey Module (HFSSM). This is accomplished by comparing model-data fit when modeling food insecurity using the one- and two-parameter logistic item response theory (IRT) models and bifactor measurement models. The reliability of adult and child subscales, based on the adult- and child-referenced items of the HFSSM, is also examined.

5.1 INTRODUCTION

The responses to the HFSSM represent two clusters of items related to household/adult and child food insecurity (Nord & Coleman-Jensen, 2014). Responses to the HFSSM may be affected because of differences in the severity of food insecurity represented by these two clusters of items. For example, Nord and Coleman-Jensen suggest a cross-classification method in which a Rasch model is estimated separately for the adult-referenced items (Items 1 – 10) and for the child-referenced items (Items 11 – 18) of the HFSSM. Coleman-Jensen, Rabbitt, and Gregory (2017) also examined methodology for determining the food security status of adults and children in the household, separately, and for the household in general. These previous studies have recognized the possible multidimensionality of the HFSSM, but there has not been a detailed and specific exploration of the dimensionality of the HFSSM based on the

[4] An earlier version of this chapter was presented at the annual meeting of the National council on Measurement in Education (Tanaka, Engelhard, Rabbitt, & Jennings, 2018).

household/adult and child items. This study specifically focuses on examining the requirement of unidimensionality using bifactor measurement models and explores whether adult and child factors appear in our data related to households with children.

The purpose of this study is to explore the use of the bifactor measurement model for evaluating the dimensionality of food security in households with children. The current instrument used to measure household food security, the HFSSM, is assumed to be unidimensional. However, this study explores the view of household food security as a multidimensional construct among households with children.

5.2 METHODOLOGY

PARTICIPANTS

The data in this study contain all households who responded in 2012 to 2015 with income below 185 percent of the federal poverty line, and who indicated having at least one child under the age of 18 in the household. The income threshold of 185 percent of the federal poverty line was used because it represents the income screening threshold for a household to be administered the HFSSM. Households above the income threshold are also administered the HFSSM if they showed any signs of food stress; however, they represented a small proportion of the households administered the HFSSM during our analysis period. Omitting higher income households from the analyses mitigates bias associated with income that could result from the screen (Nord, 2012). Households with extreme responses (scores of zero or 18) to the HFSSM were also excluded. The total number of households included in this study is 9,620.

ANALYSES

First, exploratory factor analyses were conducted using the tetrachoric correlations. These analyses indicated that three of the items on the HFSSM exhibited a high level of local dependence. The specific items deleted involved paired items that asked about the frequency of certain behaviors. Here is an example (with affirmative responses in bold):

> 9. In the last 12 months did you or other adults in your household ever not eat for a whole day because there wasn't enough money for food? (**Yes**/No)
>
> 10. (If yes to question 9) How often did this happen—**almost every month, some months but not every month**, or in only 1 or 2 months?

Items 6, 10, and 17 are related to frequencies of behaviors and were deleted from further analyses in this study. The eigenvalues for the reduced matrix of tetrachoric correlations are as follows: 8.758, 1.885, 0.901, 0.710, 0.663, 0.411, 0.373, 0.286, 0.219, 0.189, 0.160, 0.153, 0.114, 0.100, and 0.077. These eigenvalues suggest one general factor (eigenvalue = 8.758) and perhaps two secondary residual factors (eigenvalues of 1.885 and .901).

Reise (2012) suggested fitting several models in addition to the bifactor model to determine whether the bifactor model is a good choice. In this study, we fit two unidimensional models (1PL and 2PL) and two multidimensional models (correlated factors model and the bifactor model). Figure 5 contains the path diagrams for each model. Responses to the 15 items from the HFSSM were examined using these models. All models were estimated using the R package MIRT (Chalmers, 2012) and confirmed with the Mplus software (Muthén & Muthén, 2011). We only report the MIRT results here because the two programs yielded comparable results.

We examined a total of four models including two unidimensional models (the standard 1PL and 2PL models) and two multidimensional models (two bifactor models). First, we estimated a constrained bifactor model with slope parameters fixed to be equal for the general household factor, as well as fixed within each of the residual factors representing the adult and child items. Second, we estimated a bifactor model with the slopes free to vary based on the data.

5.3 RESULTS

PRELIMINARY ANALYSES

Table 10 presents the endorsement proportions for the sample of households used in this study. The endorsement proportions range from .004 [Item 18; *Child(ren) not eat for whole day*] to .839 (Item 1; *Worried food would run out*). The comparison of the four models examined in the preliminary analyses of the HFSSM are presented in Table 11. The multidimensional models have better fit than the unidimensional models. The information indices for these multidimensional models indicate that the bifactor model (AIC = 96,291.04; BIC = 96,613.76; sample-size adjusted BIC = 96,470.76) provides the best fit when compared with the correlated factors model (AIC = 97,412.19; BIC = 97,634.51; sample-size adjusted BIC = 97,536.00), justifying further examination of the bifactor model. Based on these preliminary results, we focus on the bifactor model with three factors and 15 items for the remainder of our analyses.

Additional fit statistics are given in Table 12 for the 1PL, 2PL, constrained and unconstrained bifactor models. The information criteria and likelihood ratio test reveal that the 2PL logistic model (AIC = 100,611.9; BIC = 100,827.1; sample-size adjusted BIC = 100,731.7) was a better fit to the data than the 1PL logistic model (AIC = 102,254.0; BIC = 102,368.7; sample-size adjusted BIC = 102,317.9) as expected. This result was also supported by the other fit indices (RMSEA, CLI, and TFI). The unconstrained bifactor model (AIC = 96,291.04; BIC = 96,613.76; sample-size adjusted BIC = 96,470.76) was also a better fit to the data than the constrained bifactor model (AIC = 99,365.95; BIC = 99,495.04; sample-size adjusted BIC = 99,437.84), a result that is also supported by the other fit indices.

BIFACTOR MODELS

The factor loadings for all four models are presented in Tables 13 and 14. Although we might expect all of the constrained loadings to be the same across the 15 items on the general factor, because the slopes are set to be constant there is variation related to the secondary factors that influence the estimation of the loadings resulting in variation in the factor loadings (R. P. Chalmers, personal communication, April 4, 2018). It is also important to note that the factor loadings can be interpreted as partial-correlation coefficients within the context of regression of the factors or latent variables.

When using bifactor models, it is useful to compare the general factor to the specific factors. Such comparisons identify subsets of items for which the multidimensionality might be weak enough to ignore (Stucky & Edelen, 2014). This is accomplished by calculating the Explained Common Variance (ECV), an indicator of unidimensionality. It is expressed as:

$$ECV = \frac{\sum \lambda_{Gen}^2}{(\sum \lambda_{Gen}^2) + (\sum \lambda_{Spec_k}^2)}.$$

There have been a variety of rules of thumb for judging unidimensionality. For example, Reckase (1979) suggested that values 20 percent of the variance are sufficient for creating a unidimensional scale, while Stucky and Edelen (2014) suggest values of 85 percent or higher for sets of items to be considered sufficiently unidimensional (Stucky & Edelen, 2014). As with other issues in psychometrics, the substantive conclusions related to unidimensionality depends on the structure of the domains within a scale (items groups by domains or subscales may suggest potential multidimensionality), as well as the intent and design of the developers of the scale.

The ECVs for the bifactor models in this study are shown in Table 14. In the constrained bifactor model, the ECV for the general factor was 66.1 percent, and 7.1 percent and 26.9 percent for the adult and child factors respectively. The unconstrained bifactor model had an ECV of 62.1 percent for the general factor and 23.7 percent and 14.2 percent, respectively, for the adult and child factors. These results support the inference that there is a strong general factor representing household food insecurity with some evidence for interpretable residual or specific factors. It is interesting to note that the child factor accounts for 26.9 percent of the variance under the constrained bifactor model, while the adult factor accounts for a comparable amount (23.7 percent) of variance under the unconstrained bifactor model. However, the large amount of variance accounted for by the general factor under both models supports the inference of unidimensionality for the HFSSM among households with children. This conclusion is also supported by the low reliability of the residual or specific factor scores related to the adult and child items after controlling for overall household food insecurity.

The ECV can also be calculated at the item-level. This item-level ECV indicates how well the item represents the general household factor. Low item-level ECVs (near zero) indicate a stronger association with the specific factor, while high item-level ECVs (near one) reflect a stronger association with the general household factor. For the constrained bifactor model, the average item-level ECVs were 85.5 percent for adult items and 47.6 percent for child items. For the unconstrained bifactor model, Items 3 (*Could not afford to eat balanced meals*), 15 [*Child(ren) hungry*], and 16 [*Child(ren) skipped meals*] were most strongly associated with the general household factor.

COMPARISON OF HOUSEHOLD FOOD INSECURITY ESTIMATES ACROSS MODELS

There are a variety of latent variable scores that can be estimated based on various models. This section examines estimates of latent food insecurity from four models: the two unidimensional models (1PL, 2PL) and the two multidimensional bifactor models (constrained and unconstrained, described in the previous section). The interpretation of the latent variable scores from the unidimensional models are relatively straight-forward, while the latent variable scores from the multidimensional models are more complex and open to potential misinterpretations. As pointed out by DeMars (2013), it is particularly important to recognize that the scores from bifactor models include a general latent variable (household food insecurity in this study), as well as scores on the specific factors (adult and child food insecurity) that require interpretation as residual scores after controlling for the general latent variable.

Table 15 summarizes the relationships between the latent variable scores obtained from the models. Since the Rasch model was used in the original calibration of the

HFSSM, we use the 1PL scores as our reference scale on the x-axis in order to aid in our substantive interpretations of the bifactor scores. The 1PL model differs from the standard Rasch model based on the use of a fixed slope parameter that may or may not be equal to one. The 1PL model exhibits the same invariance properties as the Rasch model (Raykov & Marcoulides, 2011). Reliability was highest for the unidimensional models.

Figure 6 shows the relationships between the 1PL model and other scoring systems. Panel A indicates the strong relationship (as expected) between the sum scores and the 1PL scores (R^2 = .9945). Since the HFSSM was developed to be a unidimensional scale based on the Rasch model, the correlations are high with the scores based on the 2PL (Panel C) and the general factors from the two bifactor models (Panels B and D). Figure 7 shows the relationships between the 1PL scores and the subscale scores obtained from the residual factors from each bifactor model. As expected, Panels A and B show weak relationships between the scores obtained based on the adult and child items (R^2 = .3924 and R^2 = .1467 respectively) with the constrained bifactor models. The relationships for the unconstrained bifactor models (Panels C and D) are also weak (R^2 = .2580 and R^2 = .1409 for adult and child respectively). Taken together, Figures 6 and 7 indicate that after accounting for the general household factor, the adult and child factors provide little additional information beyond the general factor.

DeMars (2013) suggested examining these conditional relationships to assist in the substantive interpretations of the scores. In order to illustrate her suggestion, we selected households with sum scores of 10. Figure 8 shows the relationships between estimates of general household food insecurity, and adult and child food insecurity under both the constrained and unconstrained bifactor models for households. Panels A and B indicate that the general factor obtained under the constrained bifactor model has a strong and positive correlation with the scores on the adult factor, while it has a strong and negative correlation with the child factor. That is, households that received a high score on the general factor also obtained a high score on the adult factor, but a low score on the child factor. Panels C and D indicate that the relationship between the general factor under the constrained bifactor model is negatively correlated with the adult factor and uncorrelated with the child factor. Of course, it should be noted that these relationships are conditional on the sum score of 10, and researchers should examine other sum scores when evaluating a scale.

5.4 DISCUSSION

The HFSSM is used by the USDA to measure food insecurity in the United States. It is also used to measure household-level food insecurity in other countries and has been influential in developing a global standard for measuring individual experiences of food insecurity (Cafiero, Viviani, & Nord, 2018). Previous research acknowledges the HFSSM may exhibit multidimensionality in households with children (Nord & Coleman-Jensen, 2014; Coleman-Jensen, Rabbitt, & Gregory, 2017). The purpose of this study was to examine the dimensionality of the HFSSM. Our analyses included the estimation of two unidimensional models (the 1PL and 2PL models), and two multidimensional models (a constrained and unconstrained bifactor model). The data suggest that the HFSSM is well represented as a unidimensional scale. Residual factors based on adult and child items are very unreliable, and it is not recommended that these subscale scores by reported and used based on the HFSSM for households with children.

The bifactor measurement models used in this study offer a useful way to evaluate the dimensionality of food security in households with children. The HFSSM was developed as a unidimensional scale based on Rasch measurement theory (Bickel et al., 2000; Bond & Fox, 2015). The data analyzed in this study support the conclusion that the HFSSM defines a strong single factor representing general household food insecurity. Future research is needed to replicate this finding in other populations, and to explore the potential utility of residual factors related to adult and child items in the HFSSM. The current study suggests that the reliability of the scores on these subscales are very low, and that the HFSSM yields a unidimensional scale as intended by the developers of the scale.

In Chapter 6, I discuss the implications of this study, as well as the previous two studies, for the invariant measurement of household food insecurity. The discussion is organized around the three facets of invariant measurement: household measurement, item calibration, and dimensionality.

CHAPTER 6

DISCUSSION

The purpose of this study is to examine the properties of invariant measurement in the context of household food security and the Household Food Security Survey Module (HFSSM). In this last chapter, I summarize the results of the three studies discussed in Chapters 3, 4, and 5. My summary and discussion are organized around the requirements of invariant measurement. I also identify areas for future research.

6.1 HOUSEHOLD MEASUREMENT

The first facet of invariant measurement requires item-independent measurement of households and non-crossing person response functions (PRFs; Engelhard, 2013). The household measurement study presented in Chapter 3 addresses the question of PRFs and is intended to demonstrate the value of assessing graphical displays of household fit when determining whether the requirements of invariant measurement have been met. First, preliminary analyses were completed in Facets to provide a reference point for the FDA. Then, PRFs were drawn in R using FDA techniques. The PRFs were also clustered to facilitate interpretation, an especially important step with large-scale assessment data. Misfitting households can have a negative effect on the validity of the food security status classifications made using the HFSSM. This study further illustrates the issue with relying solely on numerical indices to diagnose aberrant household response patterns and household misfit—that is, even with the same categories of misfit, there can be quite a lot of variation.

6.2 ITEM CALIBRATION

The second facet of invariant measurement describes item calibration: person-invariant calibration of items, and non-crossing item response functions (Engelhard, 2013). The study, which is presented in Chapter 4, addresses the first of these questions by examining the HFSSM for potential differential item functioning (DIF) based on participation in the Supplemental Nutrition Assistance Program (SNAP). The results of this study indicate that SNAP-based DIF does exist for the first three adult-referenced items, and the first two child-referenced items. Of these five items, three (the first adult-referenced and both child-referenced items) favored the SNAP group. That is, households that received food assistance were more likely to endorse these items than households that did not, indicating greater food insecurity. Additional analyses also showed that rates of aberrant household responses were similar in households regardless of their SNAP participation.

6.3 DIMENSIONALITY

The final facet of invariant measurement is the requirement of unidimensionality. If the principles of invariant measurement hold, then the HFSSM should be measuring only latent household food insecurity. This assumption is tested in Chapter 5, when two unidimensional (one- and two-parameter) logistic item response theory (IRT) models are compared to two multidimensional bifactor models. The bifactor models capture the general factor household food insecurity, and the specific factors adult and child food insecurity. Although the bifactor models had better fit to the data, the adult and child factors provided little additional information after accounting for the general factor. It was also found that the adult and child subscales have low reliability. For these reasons, it was concluded that the assumption of unidimensionality holds and that the unidimensional one-parameter IRT model is sufficient for modeling household food insecurity.

6.4 SUGGESTIONS FOR FUTURE RESEARCH

In summary, this dissertation explored the psychometric properties of the primary scale used to measure food insecurity nationwide. I accomplished this by applying the requirements of invariant measurement, as understood under the Rasch model, and using these requirements to guide my inquiry. The major findings can be summarized in three areas: household fit, item fit, and dimensionality. My investigation of household fit reveals FDA and clustering are useful methods for understanding food insecurity at a household level. The item fit study provided evidence that there is some bias present in key items of the scale when households participate in the food assistance program, SNAP. Finally, the dimensionality study demonstrated that though the data suggest a multidimensional model is a better fit to food insecurity data, low reliability means the unidimensional model is preferred.

Future research can expand on this work in several areas. An exciting area for additional work is in the use of FDA clustering. It would be useful to make a substantive interpretation of the four clusters identified in this dissertation. This could provide useful insight into how households in the same classification group experience food insecurity differently. Another promising topic is response patterns and aberrant response rates. Studying item response patterns could provide additional information about how households and items interact, which could, in turn, be further explored using the response functions described in Chapter 3. It would also be helpful to explore the child and adult factors identified in Chapter 5 as residual factors, and to further explore the issues of reliability of these subscales.

The aim of psychometrics is to quantify unobservable, meaningful phenomena. This is important because we want our measures to provide accurate information that we can use to inform decisions regarding policy. This dissertation investigates the scale used by the USDA, the Household Food Security Survey Module, and evaluates the psychometric properties based on the assumptions of the Rasch model.

REFERENCES

Allport, G. W. (1937). *Personality: A psychological interpretation*. New York, NY: Holt, Rinehart, & Winston.

Baker, F. B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Bickel, G., Nord, M., Price, C., Hamilton, W., & Cook, J. (2000). *Guide to measuring household food security: Revised 2000.* Washington, D.C.: U.S. Department of Agriculture, Food and Nutrition Service.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Routledge.

Cafiero, C., Viviani, S., & Nord, M. (2018). Food security measurement in a global context: The food insecurity experience scale. *Measurement, 116*, 146-152.

Carlson, A., Lino, M., Juan, W., Hanson, K., & Basiotis, P. P. (2007). *Thrifty Food Plan, 2006*. U.S. Department of Agriculture, Center for Nutrition Policy and Promotion, (2007). Retrieved from https://www.cnpp.usda.gov/sites/default/files/usda_food_plans_cost_of_food/TFP2006Report.pdf.

Carroll, J. B. (1985). Defining abilities through the person characteristic function. In E.W. Roskam (Ed.), *Measurement and Personality Assessment* (pp. 121-131). North Holland: Elsevier Science Publishers.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1-29.

Coleman-Jensen, A., Rabbitt, M. P., & Gregory, C. A. (2018). *Examining an "experimental" food security status classification method for households with children*. U.S. Department of Agriculture, Economic Research Service.

Coleman-Jensen, A., Rabbitt, M. P., Gregory, C., and Singh, A. (2019). *Household food security in the United States in 2018*. U.S. Department of Agriculture, Food and Nutrition Service.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

DeMars, C.E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing, 13,* 354-378.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.

Engelhard, G., Rabbitt, M. P., & Engelhard, E. M. (2017). Using household fit indices to examine the psychometric quality of food insecurity measure. *Educational and Psychological Measurement*, 1-19.

Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. New York, NY: Routledge.

Ferrando, P. J. (2016). Assessing person fit in typical-response measures. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 128-155). New York, NY: Routledge.

Fox, M. K., Hamilton, W., & Lin, B. H. (2004). *Effects of food assistance and nutrition programs on nutrition and health, volume 3, literature review*. U.S. Department of Agriculture, Food and Nutrition Service.

Gibson-Davis, C. M., & Foster, E. M. (2006). A cautionary tale: Using propensity scores to estimate the effect of food stamps on food insecurity. *Social Service Review*, *80*(1), 93-126.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423-436.

Gregory, C. A., Rabbitt, M. P., & Ribar, D. C. (2015). The Supplemental Nutrition Assistance Program and Food Insecurity. In J. P. Ziliak, J. Bartfield, T. Smeeding, and C. Gundersen (Eds.), *The new safety net: Why SNAP matters* (pp. 74-106). Stanford, CA: Stanford University Press.

Gundersen, C., & Oliveira, V. (2001). The food stamp program and food insufficiency. *American Agricultural Economics Association*, *83*, 875-887.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2,* 41-54.

Jennings, J. K., & Engelhard, G. (2017). Evaluation of model-data fit by comparing parametric and nonparametric item response functions: Application of a Tukey-Hann procedure. *Journal of Applied Measurement*, *18*(1), 1-13.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement, 1*(4), 477-482.

Linacre, J. M. (2012). *Winsteps tutorial 4: Differential item functioning and dimensionality*. Beaverton, OR: Winsteps.com.

Linacre, J. M. (2015). *Facets computer program for many-facet Rasch measurement, version 3.71.4*. Beaverton, OR: Winsteps.com.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Mittag, N. (2013). *A method for correcting for misreporting applied to the food stamp program*. U.S. Census Bureau Center for Economic Studies Paper Number CES-WP-13-28.

Mittag, N. (2016). *Correcting for misreporting of government benefits*. IZA Discussion Paper Number 10266.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives*, *2*(4), 201–218.

Mosier, C. I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review, 47*, 355-366.

Mosier, C. I. (1941). Psychophysics and mental test theory. II. The constant process. *Psychological Review*, *48*, 235-249.

Muthén, L. K., & Muthén, B. O. (2011). *Mplus user's guide (6th ed.).* Los Angeles, CA: Author.

Nord, M. (2012). *Assessing potential technical enhancements to the US household food security measures*. U.S. Department of Agriculture, Economic Research Service.

Nord, M., & Coleman-Jensen, A. (2014). Improving food security classification of households with children. *Journal of Hunger & Environmental Nutrition*, *9*(3), 318-333.

Nord, M., & Golla, A. M. (2009). *Does SNAP decrease food insecurity? Untangling the self-selection effect*. U.S. Department of Agriculture, Food and Nutrition Service.

Perkins, A., & Engelhard, G. (2009). Crossing person response functions. *Rasch Measurement Transactions, 23*(1), 1183-1184.

Rabbitt, M. P. (2018). Causal inference with latent variables from the Rasch model as outcomes. *Measurement*, *120*, 193-205.

Rabbitt, M. P., & Coleman-Jensen, A. (2017). Rasch analyses of the standardized Spanish translation of the US household food security survey module. *Journal of Economic and Social Measurement*, *42*(2), 171-187.

Ramsay, J. O. (1997). A functional approach to modeling test data. In Wim J. van der Linden & Ronald K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 381-394). Boca Raton, FL: CRC Press.

Ramsay, J. O. (2016). Functional approaches to modeling response data. In Wim J. van der Linden (Ed.), *Handbook of Item Response Theory, Volume One, Models* (pp. 337-350). Boca Raton, FL: CRC Press.

Ramsay, J. O. (2018). *Functional data analysis*. R package. https://cran.r-project.org/web/packages/fda/fda.pdf.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis*. (2nd ed.). New York, NY: Springer-Verlag.

Rasch (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Berkley Symposium on Mathematics, Statistics, and Probability*. Berkeley, CA: University of California Press.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. New York, NY: Routledge.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667-696.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Research Methods in Human Resource Management*, *18*(4), 210-222.

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2,* 66-78.

Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 183-206). New York, NY: Routledge.

Tanaka, V. T., & Engelhard, G. (2019). *An idiographic perspective on dimensionality using person response functions*. Electronic board presented annual meeting of the National Council on Measurement in Education in Toronto, Ontario.

Tanaka, V. T., Engelhard, G., & Rabbitt, M. P. (2019). Examining differential item functioning in the Household Food Insecurity Scale: Does participation in SNAP affect measurement invariance? *Journal of Applied Measurement*, *20*(1), 100-111.

Tanaka, V. T., Engelhard, G., Rabbitt, M. P., & Jennings, J. K. (2018). *Measuring food security using a bifactor model in households with children*. Electronic board presented at the annual meeting of the National Council on Measurement in Education in New York City, NY.

Trabin, T. E., & Weiss, D. J. (1979). *The person response curve: Fit of individuals to item response theory models*. (Research Report 79-7). Minneapolis, MN: University of Minnesota, Department of Psychology.

United States. President's Task Force on Food Assistance. (1984). Report of the President's Task Force on Food Assistance. Retrieved from http://proxy-remote.galib.uga.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edshtl&AN=mdp.39015007765806&site=eds-live

Wainer, H., Bradlow, E. T., & Wang, X. (2007). Testlet response theory and its applications. New York, NY: Cambridge University Press.

Walker, A. A. (2017). Why education practitioners and stakeholders should care about person fit in educational assessments. *Harvard Educational Review*, *87*(3), 426-444.

Walker, A.A., & Engelhard, G. (2014). Game-based assessments: A promising way to create idiographic perspectives. *Measurement: Interdisciplinary Research & Perspectives. 12*, 57–61.

Wells, C. S., & Hambleton, R. K. (2016). Model fit with residual analyses. In W. J. van der Linden (Ed.). *Handbook of Item Response Theory* (pp. 395-413). Boca Raton, FL: CRC Press.

Wilde, P., & Nord, M. (2005). The effect of food stamps on food security: A panel data approach. *Review of Agricultural Economics*, *27*(3), 425-432.

Wilson, M. (2005). *Constructing measures: An item response modeling approach* (2nd ed.). Mahwah, NJ: Erlbaum.

Wright, B. D. (1980). Afterword. In Rasch (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).

Wunderlich, G. S., & Norwood, J. L. (Eds.). (2006). *Food insecurity and hunger in the United States: An assessment of the measure*. Washington, D.C.: The National Academies Press.

APPENDIX A

TABLES

Table 1 The Household Food Security Survey Module (HFSSM)

1. “We worried whether our food would run out before we got the money to buy more.” Was that **often, sometimes,** or never true for you in the last 12 months?
2. “The food that we bought just didn’t last and we didn’t have money to get more.” Was that **often, sometimes,** or never true for you in the last 12 months?
3. “We couldn’t afford to eat balanced meals.” Was that **often, sometimes,** or never true for you in the last 12 months?
4. In the last 12 months, did you or other adults in the household ever cut the size of your meals or skip meals because there wasn’t enough money for food? (**Yes**/No)
5. (If yes to question 4) How often did this happen—**almost every month, some months but not every month,** or in only 1 or 2 months?
6. In the last 12 months, did you ever eat less than you felt you should because there wasn’t enough money for food? (**Yes**/No)
7. In the last 12 months, were you ever hungry, but didn’t eat, because there wasn’t enough money for food? (**Yes**/No)
8. In the last 12 months, did you lose weight because there wasn’t enough money for food? (**Yes**/No)
9. In the last 12 months did you or other adults in your household ever not eat for a whole day because there wasn’t enough money for food? (**Yes**/No)
10. (If yes to question 9) How often did this happen—**almost every month, some months but not every month,** or in only 1 or 2 months?

(***Questions 11-18 were asked only if the household included children age 0-17***)

11. “We relied on only a few kinds on low-cost food to feed our children because we were running out of money to buy food.” Was that **often, sometimes,** or never true for you in the last 12 months?
12. “We couldn’t feed our children a balanced meal, because we couldn’t afford that.” Was that **often, sometimes,** or never true for you in the last 12 months?
13. “The children were not eating enough because we just couldn’t afford enough food.” Was that **often, sometimes,** or never true for you in the last 12 months?
14. In the last 12 months, did you ever cut the size of any of the children’s meals because there wasn’t enough money for food? (**Yes**/No)
15. In the last 12 months, were the children ever hungry but you just couldn’t afford more food? (**Yes**/No)
16. In the last 12 months, did any of the children ever skip a meal because there wasn’t enough money for food? (**Yes**/No)

17. (If yes to question 16) How often did this happen—**almost every month, some months but not every month**, or in only 1 or 2 months?
18. In the last 12 months, did any of the children ever not eat for a whole day because there wasn't enough money for food? (**Yes**/No)

Note. Affirmative responses in bold. From Coleman-Jensen, A., Rabbitt, M. P., Gregory, C., and Singh, A. (2019). *Household food security in the United States in 2018*. U.S. Department of Agriculture, Food and Nutrition Service.

Table 2 Recommendations for improvements to the HFSSM

Concepts and definitions

Recommendation 3-1: USDA should continue to measure and monitor food insecurity regularly in a household survey. Given that hunger is a separate concept from food insecurity, USDA should undertake a program to measure hunger, which is an important potential consequence of food insecurity.

Recommendation 3-2: To measure hunger, which is an individual and not a household construct, USDA should develop measures for individuals on the basis of a structured research program, and develop and implement a modified or new data gathering mechanism.

Recommendation 3-3: USDA should examine in its research program ways to measure other potential, closely linked, consequences of food insecurity, in addition to hunger, such as feelings of deprivation and alienation, distress, and adverse family and social interaction.

Recommendation 3-4: USDA should examine alternate labels to convey the severity of food insecurity without the problems inherent in the current labels. Furthermore, USDA should explicitly state in its annual reports that the data presented in the report are estimates of prevalence of household food insecurity and not prevalence of hunger among individuals.

Survey measurement

Recommendation 4-1: USDA should determine the best way to measure frequency and duration of household food insecurity. Any revised or additional measures should be appropriately tested before implementing them in the Household Food Security Survey Module.

Recommendation 4-2: USDA should revise the wording and ordering of the questions in the Household Food Security Survey Module. Examples of possible revisions that should be considered include improvements in the consistent treatment of reference periods, reference units, and response options across questions. The revised questions should reflect modern cognitive questionnaire design principles and new data collection technology and should be tested prior to implementation.

Item response theory and food insecurity

Recommendation 5-1: USDA should consider more flexible alternatives to the dichotomous Rasch model, the latent variable model that underlies the current food insecurity classification scheme. The alternatives should reflect the types of data collected in the Food Security Supplement. Alternative models that should be formally compared include:

- Modeling ordered polytomous item responses by ordered polytomous rather than dichotomized item response functions.
- Treating items with frequency follow-up questions appropriately, for example, as a single ordered polytomous item rather than as two-independent questions.
- Allowing the item discrimination parameters to differ from item to item when indicated by relevant data.

Recommendation 5-2: USDA should undertake the following additional analyses in the development of the underlying latent variable model:

- Fitting models that allow for different latent distributions for households with children and those without children and possibly other subgroups of respondents.
- Fitting models that allow for different item parameters for households with and without children for the questions that are appropriate for all households in order to study the possibility and effects of differential item functioning.
- Studying the stability of the measurement system over time, possibly using the methods of differential item functioning.

Recommendation 5-3: To implement the underlying latent variable model that results from the recommended research, USDA should develop a new classification system that reflects the measurement error inherent in latent variable models. This can be accomplished by classifying households probabilistically along the latent scale, as opposed to the current practice of deterministically using the observed number of affirmations. Furthermore, the new classification system should be more closely tied to the content and location of food insecurity items along the latent scale.

Recommendation 5-4: USDA should study the differences between the current classification system and the new system, possibly leading to a simple approximation to the new classification system for use in surveys and field studies.

Recommendation 5-5: USDA should consider collecting data on the duration of spells of food insecurity in addition to the currently measured intensity and frequency measures. Measures of frequency and duration spells may be used independently of the latent variable measuring food insecurity.

Survey vehicles to measure food insecurity and hunger

Recommendation 6-1: USDA should continue to collaborate with the National Center for Health Statistics to use the National Health and Nutrition Examination Survey to conduct research on methods of measuring household food insecurity and individual hunger and the consequences for nutritional intake and other relevant health measures.

Recommendation 6-2: USDA should carefully review the strengths and weakness of the National Health Interview Survey in relation to the Current Population Survey in order to determine the best possible survey vehicle for the Food Security Supplement at a future date. In the meantime, the Food Security Supplement should continue to be conducted in the Current population Survey.

Recommendation 6-3: USDA should explore the feasibility of funding a one-time panel study, preferably using the Survey of Income and Program Participation, to establish the relationship between household food insecurity and individual hunger and how they co-evolve with income and health.

Note. From Wunderlich, G. S., & Norwood, J. L. (Eds.). (2006). *Food insecurity and hunger in the United States: An assessment of the measure*. Washington, D.C.: The National Academies Press. Reprinted here with permission of the publisher.

Table 3 Summary statistics for Rasch analyses

	Households	**Items**
Measure		
M	-2.63	0.00
SD	2.28	2.95
N	7324	18
Outfit		
M	0.73	0.93
SD	1.23	0.65
Infit		
M	0.99	0.98
SD	0.56	0.15
Separation statistic	2.30	44.96
Reliability of separation	0.84	> 0.99
χ^2	46888.0*	52908.2*
df	7324	17
Variance explained by Rasch model	60.91%	

* $p < .05$

Table 4 Summary of household fit statistics

			Infit MSE		Outfit MSE	
Label	**Description**	**Range of values**	**Freq**	**%**	**Freq**	**%**
A	Productive for measurement	$0.50 \leq MSE < 1.50$	3836	52.4	2253	30.8
B	Less productive for measurement, but not distorting of measures	$MSE < 0.50$	1880	25.7	4315	58.9
C	Unproductive for measurement, but not distorting of measures	$1.50 \leq MSE < 2.00$	1207	16.5	248	3.4
D	Unproductive for measurement, distorting of measures	$MSE \geq 2.00$	401	5.5	508	6.9
		Total	7324	100.0	7324	100.0

Note. MSE is the mean square error based on the Rasch model.

Table 5 Fit statistic groupings

Mean Squared Error (MSE)	**Interpretation**	**Fit Category**
$0.50 \leq MSE < 1.50$	Productive for measurement	A
$MSE < 0.50$	Less productive for measurement, but not distorting of measures	B
$1.50 \leq MSE < 2.00$	Unproductive for measurement, but not distorting of measures	C
$MSE \geq 2.00$	Unproductive for measurement, distorting of measures	D

Table 6 Rasch summary statistics

	Households	**Items**	**SNAP**
Measure			
Mean	−2.78	0.00	0.00
SD	2.25	3.06	0.00
Outfit Mean Square			
Mean	0.73	0.87	0.86
SD	1.23	0.61	0.13
Infit Mean Square			
Mean	0.99	0.97	1.01
SD	0.55	0.15	0.03
Reliability of separation	0.83	>.99	>.99
χ^2	24266.4*	27164.2*	0.0
df	3930	17	1
N	3931	18	2

* $p < .05$

Note. SNAP is the Supplemental Nutrition Assistance Program

Table 7 Summary statistics for items

Item	Label	Item Location	Standard Error	Infit MNSQ	Infit Category	Outfit MNSQ	Outfit Category
	Household items						
1	Worried food would run out	−5.49	0.05	0.99	A	1.45	A
2	Food bought would not last	−4.06	0.04	0.93	A	2.51	D
3	Could not afford to eat balanced meals	−3.52	0.04	1.18	A	1.59	C
	Adult items						
4	Adult(s) cut size or skipped meals	−1.68	0.05	0.75	A	0.58	A
5	Respondent ate less than should have	−1.89	0.05	0.82	A	0.77	A
6	Adult(s) cut size or skipped meals frequency follow up question	−0.94	0.05	0.78	A	0.54	A
7	Respondent hungry but did not eat	−0.13	0.06	0.93	A	0.62	A
8	Respondent lost weight	1.15	0.07	1.07	A	0.68	A
9	Adult(s) not eat for a whole day	1.63	0.08	0.91	A	0.41	B
10	Adult(s) not eat for a whole day frequency follow up question	2.10	0.09	0.89	A	0.33	B
	Child items						
11	Relied on low-cost foods for children	−3.23	0.04	1.33	A	1.65	C
12	Could not feed children balanced meals	−1.71	0.05	1.16	A	1.13	A
13	Child(ren) not eating enough	0.25	0.06	1.13	A	1.19	A
14	Cut size of child(ren)s meals	1.69	0.08	1.06	A	0.77	A
15	Child(ren) hungry	2.71	0.10	0.89	A	0.56	A
16	Child(ren) skipped meals	3.45	0.13	0.86	A	0.58	A
17	Child(ren) skipped meals frequency follow up question	3.85	0.14	0.84	A	0.20	B

18	Child(ren) not eat for a whole day	5.80	0.28	0.99	A	0.08	B

Note. MNSQ is the mean square error. The fit categories come from guidelines suggested by Engelhard and Wind (2018) for interpreting fit statistics. Category A represents $0.50 \leq$ MNSQ < 1.50. Category B represents MNSQ < 0.50. Category C represents $1.50 \leq$ MNSQ < 2.00. Category D represents MNSQ ≥ 2.00

Table 8 Supplemental Nutrition Assistance Program (SNAP) Differential Item Functioning (DIF) analysis

		SNAP Receipt		No SNAP Receipt		Difference between SNAP and No SNAP Receipt		
Item	**Label**	**Rasch measure**	***SE***	**Rasch Measure**	***SE***	**Contrast**	***t*(*df* = 3929)**	***p*-value**
	Household Items							
1	Worried food would run out	−5.81	0.07	−5.19	0.07	−0.62	−6.26	.0000*
2	Food bought would not last	−4.20	0.06	−3.89	0.06	−0.31	−3.61	.0003*
3	Could not afford to eat balanced meals	−3.28	0.06	−3.82	0.06	0.54	6.29	.0000*
	Adult Items							
4	Adult(s) cut size or skipped meals	−1.71	0.06	−1.63	0.07	−0.08	−0.89	.3752
5	Respondent ate less than should have	−1.86	0.06	−1.92	0.07	0.05	0.56	.5776
6	Adult(s) cut size or skipped meals frequency follow up question	−0.94	0.06	−0.93	0.08	−0.02	−0.17	.8644
7	Respondent hungry but did not eat	−0.11	0.07	−0.14	0.09	0.02	0.21	.8364
8	Respondent lost weight	1.06	0.08	1.34	0.12	−0.28	−1.85	.0648
9	Adult(s) not eat for a whole day	1.58	0.09	1.75	0.14	−0.16	−0.97	.3307
10	Adult(s) not eat for a whole day frequency follow up question	2.05	0.11	2.21	0.15	−0.16	−0.88	.3783
	Child Items							
11	Relied on low-cost	−3.14	0.06	−3.33	0.07	0.19	2.18	.0295*

	foods for children							
12	Could not feed children balanced meals	−1.58	0.06	−1.90	0.07	0.31	3.30	.0010*
13	Child(ren) not eating enough	0.22	0.07	0.31	0.10	−0.09	−0.72	.4723
14	Cut size of child(ren)s meals	1.69	0.10	1.71	0.14	−0.02	−0.17	.9274
15	Child(ren) hungry	2.76	0.13	2.61	0.17	0.15	0.68	.4943
16	Child(ren) skipped meals	3.49	0.16	3.38	0.21	0.11	0.40	.6896
17	Child(ren) skipped meals frequency follow up question	3.89	0.18	3.80	0.24	0.09	0.29	.7741
18	Child(ren) not eat for a whole day	5.82	0.37	5.79	0.43	0.03	0.04	.9646

Table 9 Household fit mean square error categories

	MNSQ > 1.50	MNSQ ≥ 2.00	N
Infit			
SNAP Receipt	0.231	0.047	2257
No SNAP Receipt	0.228	0.054	1674
Total	0.230	0.050	3931
Outfit			
SNAP Receipt	0.096	0.067	2257
No SNAP Receipt	0.100	0.071	1674
Total	0.098	0.069	3931

Note. Cell entries are proportions. MNSQ is the mean square error.

Table 10 Item-severity parameters (endorsements)

Item Index	Label	Endorsement Proportions
	Household/Adult items	
1	Worried food would run out	.839
2	Food bought would not last	.676
3	Could not afford to eat balanced meals	.572
4	Adult(s) cut size or skipped meals	.343
5	Respondent ate less than felt should have	.346
7	Respondent hungry but did not eat	.189
8	Respondent lost weight	.105
9	Adult(s) did not eat for whole day	.066
	Child items	
11	Relied on low-cost foods for child(ren)	.550
12	Could not feed child(ren) balanced meals	.349
13	Child(ren) not eating enough	.154
14	Cut size of child(ren)s meals	.078
15	Child(ren) hungry	.048
16	Child(ren) skipped meals	.026
18	Child(ren) not eat for whole day	.004
N = 9,620		

Note. Items 6, 10, and 17 from the full 18-item scale have been deleted due to local dependence.

Table 11 Comparison of four models for examining dimensionality of the HFSSM

Models	**Free Parameters**	**Akaike (AIC)**	**Bayesian (BIC)**	**Sample-size adjusted BIC**
Unidimensional				
1PL	17	102254.0	102368.7	102317.9
2PL	31	100612.1	100827.1	100731.7
Multidimensional				
Correlated Factors	33	97412.19	97634.51	97536.00
Bifactor	48	96291.04	96613.76	96470.76

Note. Both 1PL (item slopes are equal to a constant) and 2PL (item slopes vary) models are unidimensional.

Table 12 Summary of fit statistics (N = 9,620)

	Unidimensional		Multidimensional	
	1PL model	**2PL model**	**Constrained Bifactor model**	**Unconstrained Bifactor model**
Free Parameters	17	31	19	48
Loglikelihood	-51110.00	-50275.96	-49664.98	-48100.52
-2 LL Difference	1670.06*		3128.915*	
AIC	102254.0	100611.9	99365.95	96291.04
BIC	102368.7	100827.1	99495.04	96613.76
Sample-size adjusted BIC	102317.9	100731.7	99437.84	96470.76
RMSEA	.094	.090	.067	.036
CFI	.882	.906	.940	.987
TLI	.881	.891	.939	.982

* $p < .05$

Note. LL is the loglikelihood, AIC is the Akaike Information Criterion, BIC is the Bayesian Information Criterion, RMSEA is the Root Mean Square Error of Approximation, CFI is the Comparative Fit Index, and TLI is the Tucker Lewis Index. See the text for a description of the models.

Table 13 Factor loadings from the unidimensional models of household food insecurity

	1PL model	**2PL model**
Item Index	**Household**	**Household**
Household/Adult items		
1	0.736	0.551
2	0.736	0.651
3	0.736	0.596
4	0.736	0.867
5	0.736	0.892
7	0.736	0.879
8	0.736	0.816
9	0.736	0.808
Child items		
11	0.736	0.533
12	0.736	0.682
13	0.736	0.752
14	0.736	0.797
15	0.736	0.885
16	0.736	0.863
18	0.736	0.816

Table 14 Factor loadings from the multidimensional models of household food insecurity

	Constrained Bifactor model				Unconstrained Bifactor model			
Item Index	Household	Adult	Child	I-ECV	Household	Adult	Child	I-ECV
Household and adult items								
1	0.724	0.298	—	.855	0.273	0.682	—	.138
2	0.724	0.298	—	.855	0.502	0.443	—	.562
3	0.724	0.298	—	.855	0.841	-0.108[1]	—	.984
4	0.724	0.298	—	.855	0.657	0.622	—	.527
5	0.724	0.298	—	.855	0.675	0.643	—	.524
7	0.724	0.298	—	.855	0.669	0.626	—	.533
8	0.724	0.298	—	.855	0.608	0.623	—	.488
9	0.724	0.298	—	.855	0.606	0.613	—	.494
Child items								
11	0.593	—	0.622	.476	0.518	—	0.632	.402
12	0.593	—	0.622	.476	0.775	—	0.453	.745
13	0.593	—	0.622	.476	0.715	—	0.565	.616
14	0.593	—	0.622	.476	0.724	—	0.463	.710
15	0.593	—	0.622	.476	0.842	—	0.319	.874
16	0.593	—	0.622	.476	0.798	—	0.389	.808
18	0.593	—	0.622	.476	0.729	—	0.424	.747
ECV	.661	.071	.269		.621	.237	.142	

[1] This value is not statistically significant

Table 15 Summary of score estimates from different models

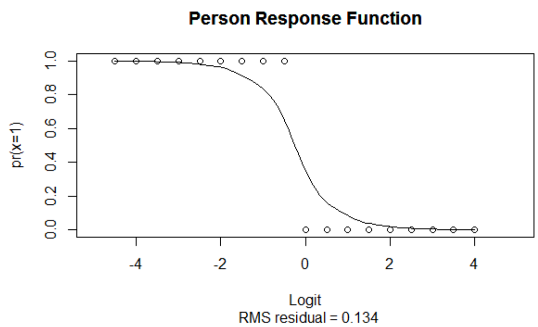
	Theta			Reliability	
Models	**Mean**	**SD**	**Mean SE**	**Empirical**	**Theoretical (SD = 1)**
Unidimensional					
1PL	0.00	0.90	0.42	0.78	0.82
2PL	0.00	0.90	0.43	0.77	0.82
Multidimensional					
Constrained Bifactor					
Household	0.00	0.85	0.53	0.61	0.72
Adult	0.00	0.38	0.92	< 0	0.15
Child	0.00	0.70	0.72	< 0	0.49
Unconstrained Bifactor					
Household	0.00	0.84	0.53	0.60	0.72
Adult	0.00	0.75	0.66	0.24	0.57
Child	0.00	0.62	0.78	< 0	0.39

Note. Theoretical values for reliability coefficients are based on assumption that latent variable has a standard deviation of one. Empirical values are estimated as $1\text{-}SE^2/SD^2$ (Wainer, Bradlow, and Wang, 2007, p. 76)

APPENDIX B

FIGURES

Panel A: [1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0]



Panel B: [1,1,1,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0]

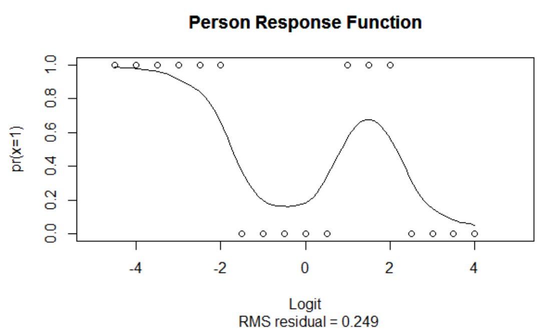


Figure 1. Person response functions estimated with FDA based on item locations for the HFSSM items (18 items).

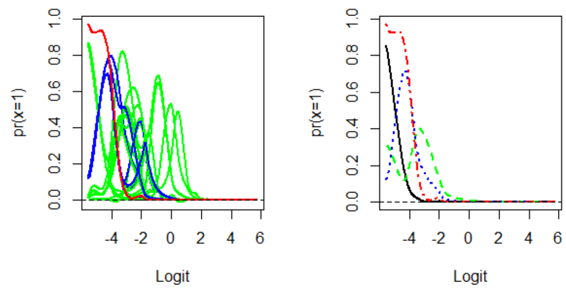
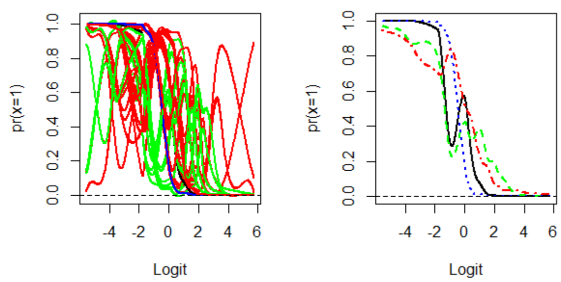
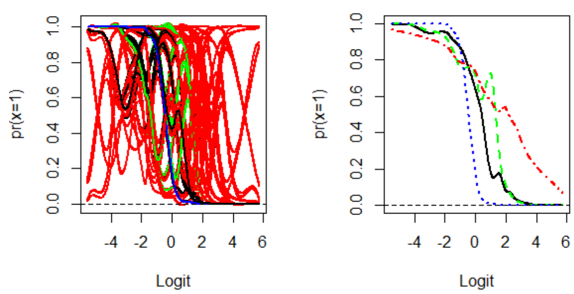
Scores	Clusters (4)
Food Secure 1-2	
Low Food Security 3-7	
Very Low Food Security 8-17	

Figure 2. Cluster analyses (four clusters) by food insecurity category. Person response functions (PRFs) were drawn using functional data analysis (FDA).

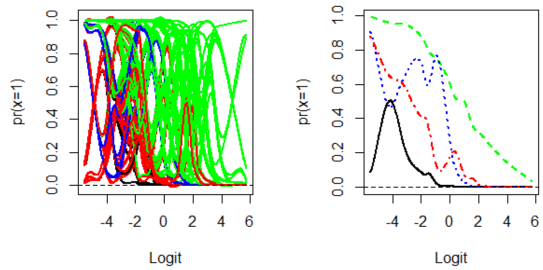
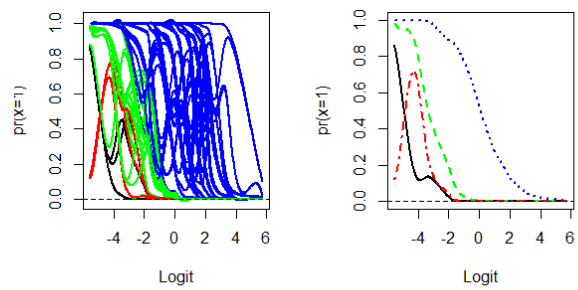
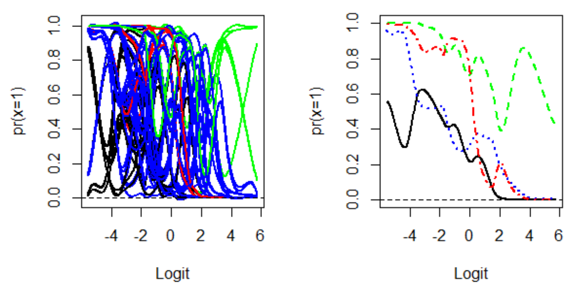
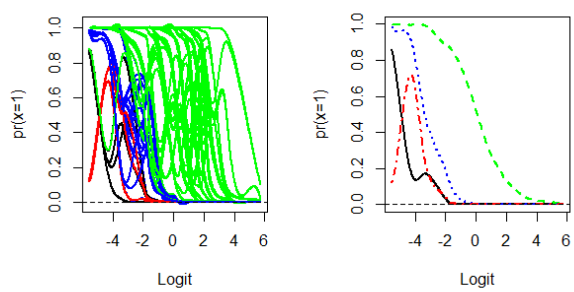
	Clusters (4)
Label	
A	
B	
C	
D	

Figure 3. Cluster analyses (four clusters) by fit grouping. Person response functions (PRFs) were drawn using functional data analysis (FDA).

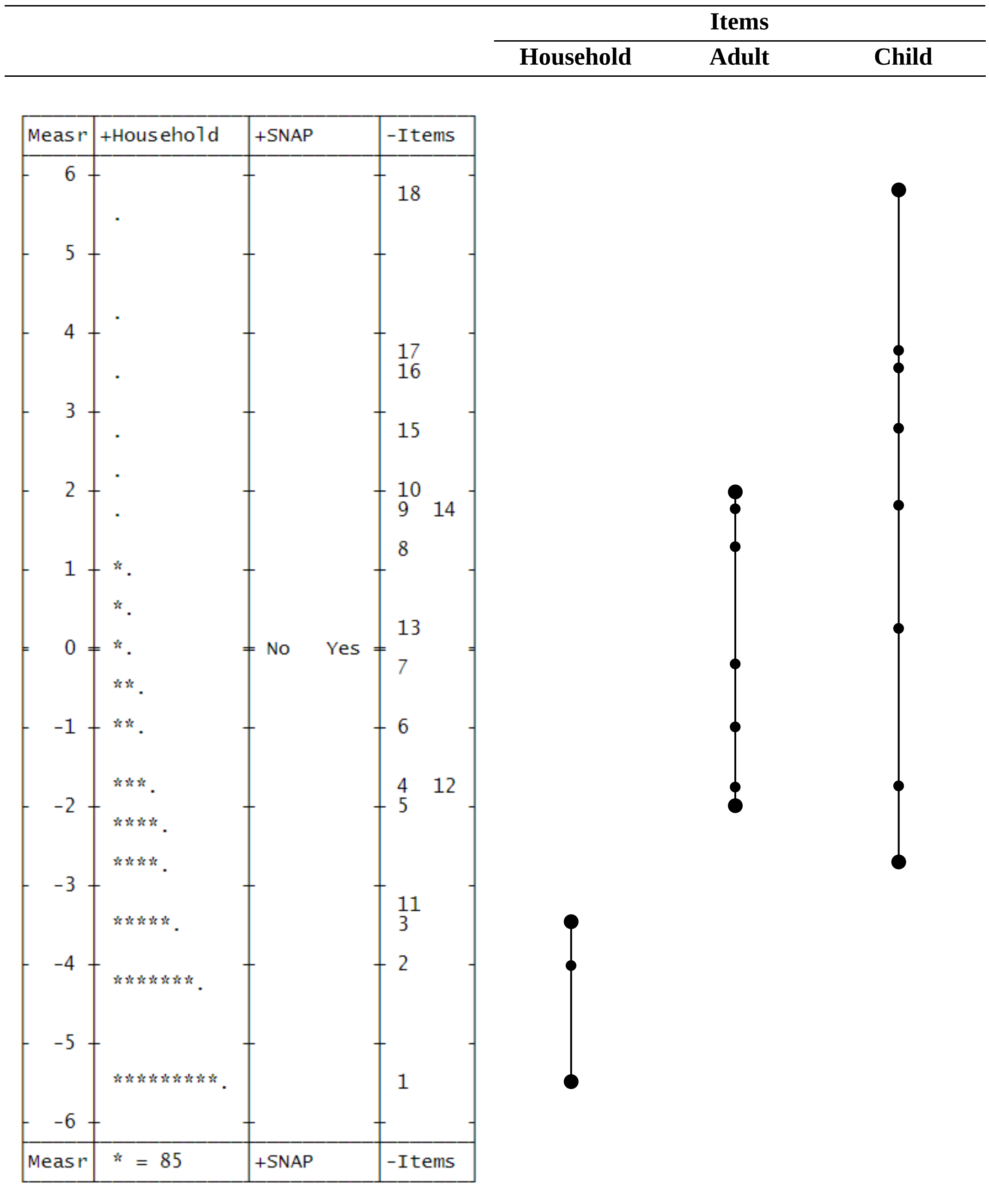


Figure 4. Wright map of the household food insecurity scale.

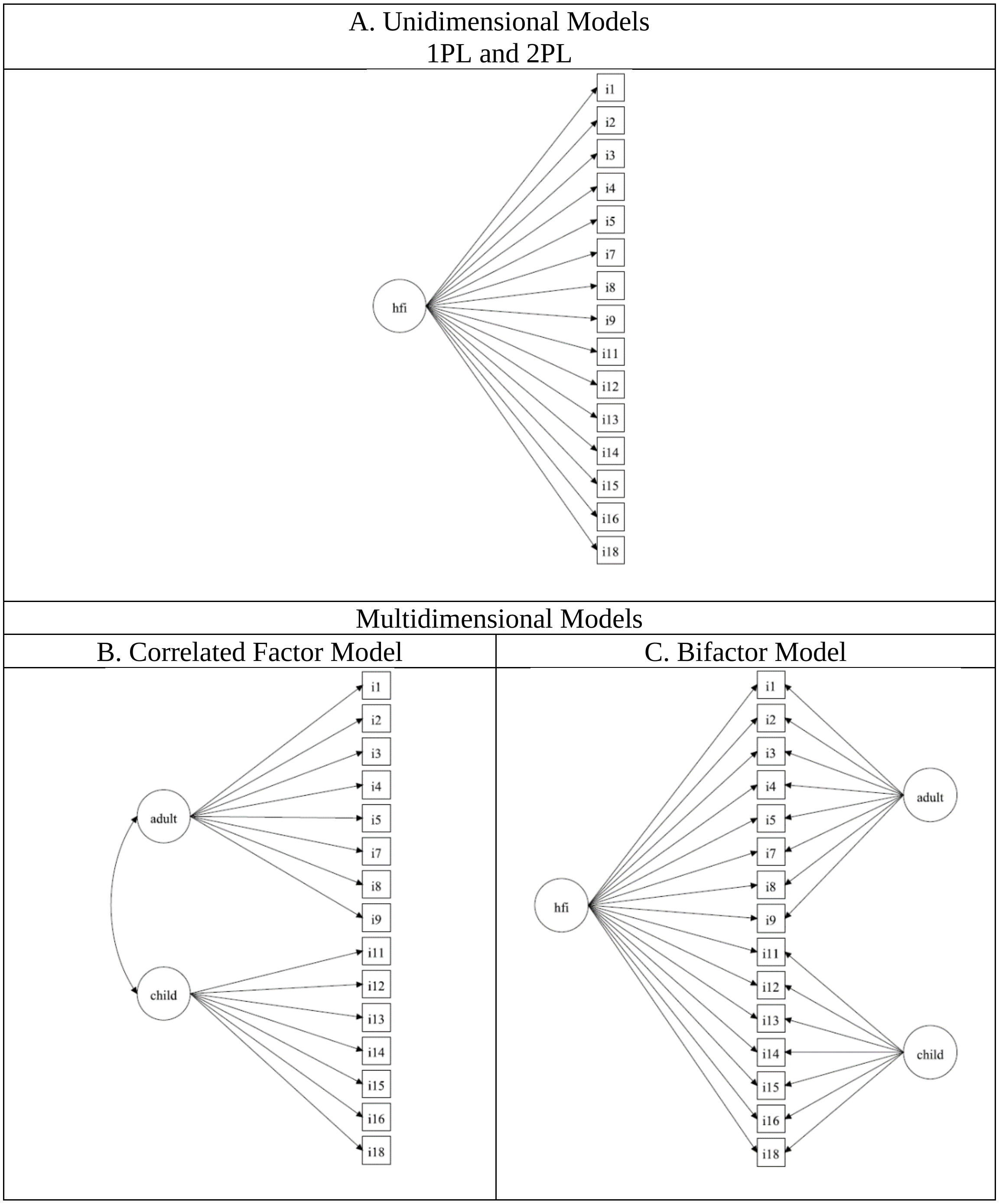
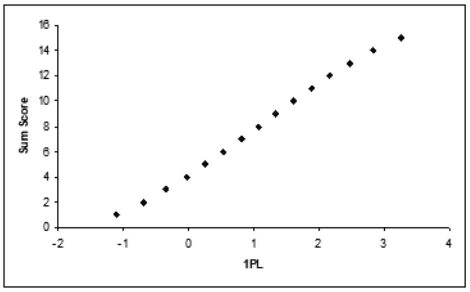
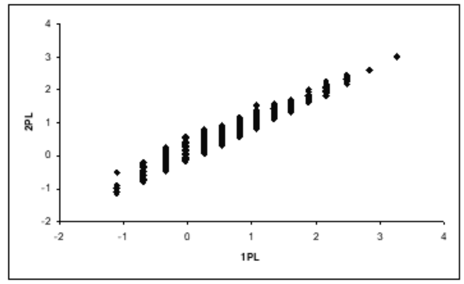


Figure 5. Path diagrams for the unidimensional and multidimensional analyses of the 15-item Household Food Security Survey Module (HFSSM).

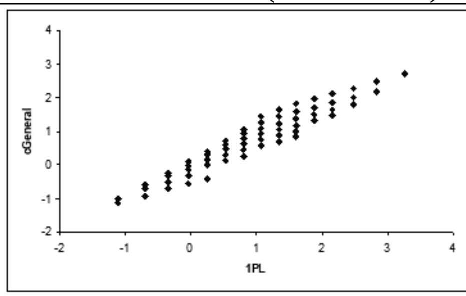
A. Sum Score on 1PL (R^2 = 0.9945)



C. 2PL on 1PL (R^2 = 0.9719)



B. cGeneral on 1PL (R^2 = 0.9794)



D. uGeneral on 1PL (R^2 = 0.8710)

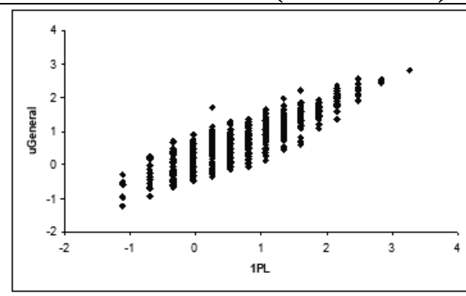
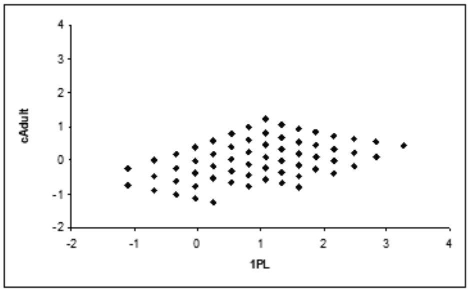
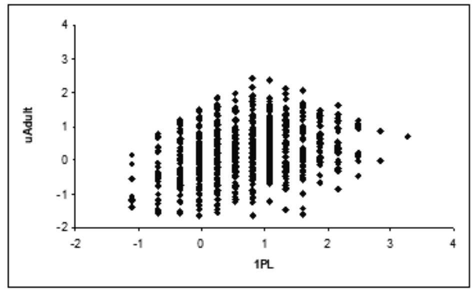


Figure 6. Relationships between 1PL scores and other scoring systems for general factor.

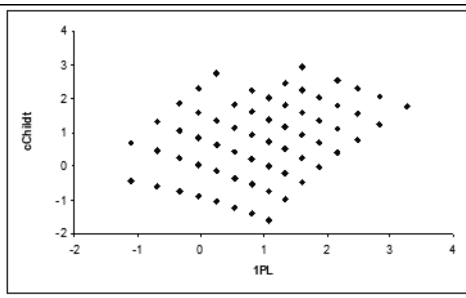
A. cAdult on 1PL ($R^2 = 0.3924$)



C. uAdult on 1PL ($R^2 = 0.2580$)



B. cChild on 1PL ($R^2 = 0.1467$)



D. uChild on 1PL ($R^2 = 0.1409$)

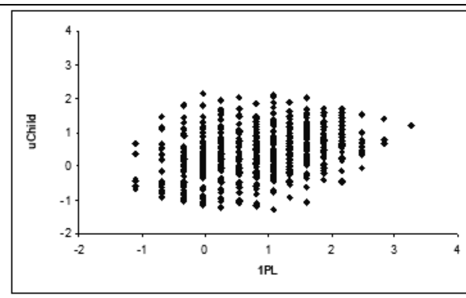
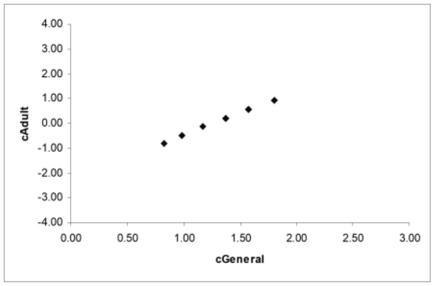
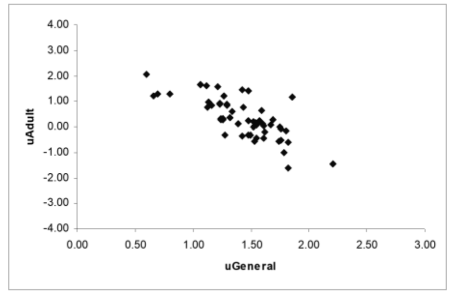


Figure 7. Relationships between 1PL scores and other scoring systems for specific factors.

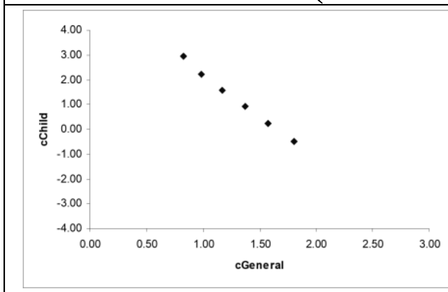
A. cAdult on cGeneral ($R^2 = 0.9994$)



C. uAdult on uGeneral ($R^2 = 0.5483$)



B. cChild on cGeneral ($R^2 = 0.9991$)



D. uChild on uGeneral ($R^2 = 0.0527$)

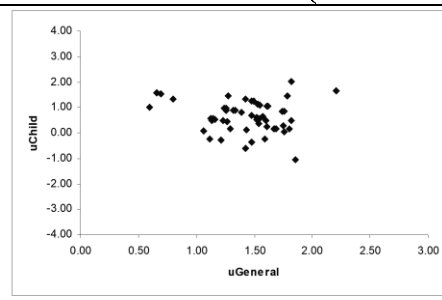


Figure 8. Relationships between the constrained and unconstrained factors for households with a sum score of 10. cAdult is the constrained adult factor, cGeneral is the constrained general factor, cChild is the constrained child factor, uAdult is the unconstrained adult factor, uGeneral is the unconstrained general factor, and uChild is the unconstrained child factor.