

# Methodological issues in measuring food insecurity and hunger

*Matthew S. Johnson*<sup>1</sup>

## INTRODUCTION

The USDA's Food Security program's goal is to estimate the number of people in the United States who are food secure, and food insecure with or without hungry. In order to estimate the number of individuals in the country who are food insecure, we must first be able to classify individuals into one of the food security levels, or at least be able to find the probabilities of class membership for each individual. The Life Sciences Research Office provides the following definitions of food insecurity and hunger (Anderson, 1990):

**Food Insecurity**—Limited or uncertain availability of nutritionally adequate and safe foods or the uncertain ability to acquire acceptable foods in socially acceptable ways.

**Hunger**—The uneasy or painful sensation caused by a lack of food; or, the recurrent and involuntary lack of access to food. Hunger is a potential, although not necessary, consequence of food insecurity.

Both definitions above describe directly observable phenomenon. If we could follow an individual for an entire year we could determine how often that person was unable to acquire acceptable foods. Obviously it would be cost prohibitive to perform such a study, thus necessitating an instrument, such as a measurement scale, by which we can indirectly measure food insecurity and possibly hunger.

This paper discusses the issues relevant to the development of a food insecurity measurement scale. The goal of the measurement scale is to develop a set of items and a scaling model by which each individual's food insecurity level, or *propensity*,  $\theta_i$  can be measured. The estimated propensities could then be used to estimate the proportion of the population that falls into each food insecurity class defined by cutpoints on the propensity scale. One point of contention in the measurement of food insecurity and hunger has been whether hunger is truly an extreme case of

---

<sup>1</sup>Matthew Johnson is Assistant Professor of Statistics in the Department of Statistics & Computer Information Systems at Baruch College of the City University of New York, New York, NY 10010.

food insecurity. Rather than enter into that discussion, this paper will assume that there are different levels of food insecurity by which individuals can be classified, whether one of those classes is called “hunger” will be left to the experts in the field.

DeVellis (1991) describes the development of a measurement scale as an eight step process. Section 2 summarizes DeVellis’s scale development steps and relates them to the development of a food insecurity scale. Section 3 reviews item response theory models, and Section 4 discusses how to evaluate these models. Section 5 offers some thoughts as to how the IRT models can be utilized to define food security classes along the measurement scale, and to estimate the portion of the population within each class. Section 6 reviews techniques for maintaining the measurement scale over time.

## **DEVELLIS’S EIGHT STEPS FOR SCALE DEVELOPMENT**

### **STEP 1: Determine clearly what it is you want to measure.**

DeVellis (1991) suggests two ways to clarify what is being measured. The first is to ground the construct being measured in substantive theory. The second is to decide on the level of specificity or generality at which the construct will be measured.

The Panel to Review the U.S. Department of Agriculture’s Measurement of Food Insecurity and Hunger (henceforth referred to as the Panel) notes that the Food Insecurity definition really describes two related, but separate concepts: (National Research Council, 2005, pp 3-1---3-2)

*Uncertainty* about being able to obtain food in socially acceptable ways due to lack of resources, causing worry and mental, emotional, and physical stress.

*Insufficiency* in (or lack of access to) the quantity and quality of nutritionally adequate and safe foods.

The Panel suggests that some of the questions currently used in the Food Security survey measure uncertainty, some measure insufficiency, and others measure hunger. However, all items are treated equally in the analysis stage.

The task of the scale developers is to determine whether they would like to measure the specific constructs of “food uncertainty,” “food insufficiency,” and “hunger” separately, or whether one general construct, such as the “food insecurity” construct is adequate. It is often easier to design a measurement instrument for specifically defined constructs. A construct like food insecurity, which is multi-faceted, can be problematic when scaled, because the scale can be influenced by the number of items addressing each of the subconstructs. For example, a food security scale that contains mostly questions about uncertainty may be quite different from a scale containing mostly items about insufficiency.

### **STEP 2: Generate an item pool**

Once the construct or constructs of interest have been clearly defined the scale developer should develop a vast pool of items. Ideally, several items would be generated, many more than will be used in the final scale; and the items would measure various levels of the scale over the

range of interest. For example, the food security program is not necessarily interested in accurately measuring the food security level of food secure households, and therefore, there is no need to produce items measuring different levels of those food secure households.

In most situations it will be much easier to find ten good questions measuring food insecurity from a pool of forty, then it would be to simply come up with the best ten items immediately. So, at this stage of the development it is probably better to be “overinclusive” (DeVellis, 1991). It is quite acceptable to have two items asking about nearly the same manifestation of the construct, as long as they are asked in different ways. If the information being expressed by the items is truly redundant, then only one of the items should appear on the final questionnaire, and if no other criterion is available experts can choose the one that is easiest for the respondent to understand.

The creation of an item pool must take the readability of the items into consideration. DeVellis suggests the following guidelines when writing items:

1. Avoid exceptionally length items.
2. Choose an appropriate reading difficulty level. DeVellis suggests a reading level between the fifth and seventh grades for most scales.
3. Avoid ambiguity.
4. Avoid double-barreled items where the item could be affirmed or denied for multiple reasons.
5. Follow conventional rules of grammar.

The final consideration when developing items is whether to use positively worded items, negatively worded items, or both. The items in the current Food Insecurity scale all ask items in a direction where an affirmative response is evidence towards food insecurity, e.g., “We couldn’t afford to eat balanced meals.” Some authors have found that there can be an acquiescence bias when all items are asked in the same direction. These authors suggest reversing the directions of some of the items. The balanced meal item could be transformed to state, “We could always afford to eat balanced meals.”

Because the responses of negatively worded items can easily be reversed, the remainder of this paper will assume that all items have been asked in the same direction, e.g., all items provide evidence for food insecurity.

### **STEP 3: Determine the format for measurement.**

When developing a measurement scale the researcher must determine whether dichotomous or polytomous response formats would provide more information. If polytomous formats are going to be used the researcher must decide whether it makes sense to include a middle “neutral” category. Depending on how the survey is administered a continuous response scale might be appropriate.

The current Food Insecurity items are a mix of dichotomous and polytomous items, but all are dichotomized in analysis. Often times it is advantageous to utilize polytomously scored

items, because they typically provide more information about the latent construct being studied. However, when many response categories are utilized the respondents may be unable to distinguish between adjacent categories.

Most measurement scales utilize between four and seven response categories. The polytomous items utilized in the current Food Insecurity measurement scale all contain three response categories. The three response categories make sense for the way the questions are asked, because they are trying to get at frequency. For example, one item states, “We couldn’t afford to eat balanced meals.” The respondents are asked whether that was *often*, *sometimes*, or *never* true over the last twelve months. It might be difficult to expand the number of response categories if they need to reflect frequency.

If frequency was not an issue the items could be modified into a *Likert scale* format. The Likert format presents respondents with a declarative sentence, and then asks the respondent the level at which they agree or disagree with the statement, e.g., “Strongly Disagree”, “Disagree”, “Agree”, “Strongly Agree”. The statements should be relatively strong statements, because weak statements are often affirmed for individuals across the entire scale. For example, good Likert items might state, “We rarely could afford to eat balanced meals,” or “We often couldn’t afford to eat balanced meals”; the response categories have clear meanings with regard to these statements. The statement “There were times when we couldn’t afford to eat balanced meals,” on the other hand, is not as strong a statement. Households with moderate food insecurity could be as likely to “Strongly Agree” with the statement as households with severe food insecurity.

#### **STEP 4: Have experts review the initial item pool.**

It is always good to have one’s work reviewed by experts outside of the study. External reviewers may be able to think about the items in ways that have not occurred to the scale developers. It is also wise to employ a wide variety of experts, with different views on food insecurity. If the development team only asks for comments from researchers with their same views, then they may not get the most valuable feedback.

#### **STEP 5: Consider inclusion of validity items.**

Ideally we would be able to measure food insecurity and hunger perfectly for at least some subset of the entire sample. Being able to include this “predictive validity” information, even for a small subset of individuals, would have a very strong impact on determining the value of the scale items. Furthermore, prevalence estimates could be improved if such predictive validity information could be included.

If other scales have been developed to measure food insecure, or related constructs, then it would be beneficial to include those items in the pool. Even though you may not plan to use the items on the final measurement scale, it is worth including them in development sample discussed in STEP 6 below. In the development of the current food security measurement scale there were validity items relating to income, weekly food expenditures, and a single-item measure of food insecurity that had been used in previous research.

### **STEP 6: Administer items to a development sample.**

Any new items developed for the measurement of food insecurity need to be tested before being operationalized. Without pretesting items it will be difficult to select a small number to include in the final survey form. It is important to make sure that the development sample be representative of the operational samples on which you will be administering the final measurement scale. If the developmental sample is not inclusive enough you will not know what to expect when the scale is operationalized.

Sometimes it is useful to ask the respondents in a development sample what they are thinking about as they read and respond to each question. The information can be analyzed qualitatively to determine if the respondents are actually interpreting the questions as meant by the item writers, or if respondents are confused about what is being asked. This qualitative information can be every bit as useful as the quantitative information calculated in STEP 7 below.

### **STEP 7: Evaluate the items.**

Once development data have been collected, all new items need to be evaluated. Item analyses can be performed to determine whether items appear to be measuring the same latent construct (internal reliability). If validity items are included in the development sample data then we need to determine whether the scale items are correlated with the validity items. Depending on what model is used for analysis (e.g., item response models, factor analysis) the researcher must be sure that the assumptions of the model are reasonable for the analysis of the developed items. Some statistics that prove useful for performing an exploratory analysis of the response data are summarized below.

- *Average scores.* Items on which all respondents tend to produce a high or low score are rarely informative when trying to discriminate between the respondents.
- *Item-score correlations.* Another simple measure is to calculate the correlation between the score on a single item and the score on the remainder of the items in the measurement scale. Items with negative correlations either need to be reversed or thrown out. Items that have low (but positive) correlations tend to be poor discriminators, and it may be wise to leave them out of the final scale.

Item-score correlations can also provide some information about what model might be best for the measurement data. The item-score correlations for item responses generated from the Rasch model tend to be relatively constant, and so if the plan is to use the Rasch model in analysis, then it is important at this stage to find items in the pool with similar item-score correlations.

- *Reliability measures.* A number of measures have been proposed to examine the internal reliability of a scale, including Cronbach's alpha, and Spearman-Brown's split-half measure. Three measures were examined when developing the current food security scale. Each measure indicated that the 12-month measurement scale was fairly reliable. The 30-day scales were less reliable.
- If validity items were included in the developmental sample, the new items could be compared and contrasted to these items using the information from pre-testing. If predic-

tive validity questions were included, that information could be utilized to set cutpoints on the final measurement scale.

- *Differential item functioning.* One of the assumptions that is implicit in IRT models is that items behave similarly for all individuals with the same propensity score  $\theta$ . For example, a man and a woman with the same propensity score  $\theta$ , should produce a positive response to a question with the same probabilities. If the response behaviors of men and women are not the same, then any estimates, such as the prevalence of food insecurity, based on an IRT model will likely be biased.

There is a relatively simple analysis that can be performed to examine differential item functioning (Holland and Thayer, 1986). The method uses the raw scores  $S_i$  of the survey participants as proxies for their propensities  $\theta_i$ . A Mantel-Haenszel chi-squared test is performed to determine whether or not the score on a given item is conditionally independent of the group membership (e.g. race, gender, households with children versus those without) of the responding individual given the raw score.

Opsomer, Jensen, and Pan (2002) examine the question of differential item functioning by fitting a mixed effects model that has interactions between demographic variables and the item difficulties. They find that some of these interactions are in fact statistically significant.

- *Local dependence and dimensionality.* Two of the core assumptions in item response theory are that: (a) the items measure a single latent variable; and (b) conditional on that latent variable the responses to the items are independent. If the survey items are actually measuring a multidimensional latent trait, but a unidimensional model is fit to the data, the resulting scale is difficult, if not impossible to interpret.

Several authors have suggested methods for detecting multidimensionality from multiple item responses. Two popular procedures for examining dimensionality in IRT are the DETECT procedure (Zhang and Stout, 1999), which attempts to determine the number of dimensions in a test, and the DIMTEST procedure (Stout, 1987), which is used to test specific hypotheses about the dimensionality.

Froelich (2002) performs a dimensionality study of the 1995 food security survey data and finds that at least two dimensions are present in the item responses for households with children. One dimension of the latent construct is associated with the household/individual items, and the other dimension is associated with the questions asking specifically about children.

The researcher should either remove any items that appear problematic from consideration for the final measurement scale, or plan on modeling the phenomenon. For example, if the dimensionality assessment indicates there is more than one dimension, the researcher should be ready to deal with the multidimensionality in the analysis stage.

### **STEP 8: Optimize scale length.**

The Food Insecurity program differs from most measurement problems in that it is not neces-

sarily interested in accurately measuring food insecurity at the individual level. The goal is to get an accurate estimate of the number of food insecure individuals in the population. Therefore, it may be possible to get by with a shorter scale. Furthermore, because the goal is to estimate prevalence (and possibly frequency and duration) it is not necessary to populate the entire scale with items. Items focused near the cutpoints will provide the most information about the prevalence of each food insecurity classification.

## ITEM RESPONSE THEORY AND OTHER LATENT VARIABLE MODELS

Item response theory (IRT) models are a class of statistical models used by researchers to describe the response behaviors of individuals to a set of categorically scored items. The most common IRT models can be classified as generalized linear fixed- and/or mixed-effect models. Although IRT models appear most often in the educational testing literature, researchers in other fields have successfully utilized IRT-like models in a wide variety of applications.

To formalize the item response problem, let  $X_{ij}$  be the (possibly polytomous) score of individual  $i \in \{1, \dots, N\}$  to item  $j \in \{1, \dots, J\}$ . Further let  $P_{jm}(\theta_i) \equiv Pr\{X_{ij} = m \mid \theta_i\}$ , denote the  $m$ th *category response function* for item  $j$ . When item  $j$  is dichotomous the function  $P_j(\theta) = P_{j1}(\theta)$  is called the *item response function* (IRF) for item  $j$ .

A number of item response models exist in the statistics and psychometric literature for the analysis of multiple discrete responses. The models typically rely on the following assumptions:

- *Unidimensionality (U)*: There is a one-dimensional, unknown quantity associated with each respondent in the sample that describes the individuals propensity to endorse the items in the survey (or exam). Let  $\theta_i$  denote the propensity of individual  $i$ .
- *Conditional Independence (CI)*: Given an individual's propensity  $\theta$ , the elements of the item response vector for respondent  $i$ ,  $\mathbf{X}_i = (X_{i1}, \dots, X_{iJ})^t$ , are independent.
- *Monotonicity (M)*:  $Pr\{X_{ij} > t \mid \theta_i\}$  is a non-decreasing function of an individual's propensity  $\theta_i$ , for all  $j$  and all  $t$ . Respondents with high propensities are more likely to endorse items than those with low propensities.

In educational testing psychometricians often refer to the propensity  $\theta_i$  as the latent ability, or proficiency of individual  $i$ . In the *Food Security* program the propensity is assumed to be a measure of the food insecurity of the individual (or household).

In addition to the assumptions above, mixed-effect models assume some distribution  $F(\theta)$  for the latent variable  $\theta$ . The distribution may or may not have support on the whole real line. When the distribution  $F(\cdot)$  is discrete, we typically call the resulting model an *ordered latent class model*. *Latent variable models* usually refer to models where  $F(\cdot)$  is continuous.

In many applications that utilize measurement scales, the latent variables  $\theta_i$  are the parameters of interest. However, the Food Insecurity program is interested in how the food insecurity propensities are distributed in the population, and therefore, is interested in the shape of the distribution  $F(\theta)$ .

### Models for dichotomous item responses

The simplest type of item response model is concerned with the analysis of dichotomously scored items (correct/incorrect). In the dichotomous case, the monotonicity assumption (M) states that the *item response function (IRF)*  $P_j(\theta) \equiv Pr\{X_{ij} = 1 | \theta\}$  is a non-decreasing function of  $\theta$  for all items  $j$ .

The monotonicity assumption (M) allows us to use the observed item response vector for individual  $i$  ( $\mathbf{X}_i$ ) as repeated measures of the latent variable  $\theta$ . In fact, in the dichotomous case, under the conditions U, CI and M the *total score* for individual  $i$ , defined as  $S_i = \sum_{j=1}^J X_{ij}$  stochastically orders the propensity  $\theta$ , i.e.,

$$Pr\{\theta > t | S = s\} > Pr\{\theta > t | S = r\} \text{ for all } t \text{ and } s > r.$$

and the score  $S_i$  consistently orders individuals by their latent variable  $\theta$ .

Typically a link function is assumed that relates the propensities of the survey respondents and properties of the items to the item response function  $P_j(\theta)$ , or item-category response functions  $P_{jm}(\theta)$ . The most common link functions utilized in IRT are the probit link function (i.e., the inverse of the normal cumulative distribution function) and the logistic link function:

$$\log \left\{ \frac{P_j(\theta)}{1 - P_j(\theta)} \right\} \quad (1)$$

In the sections below I will review some common IRT models for both dichotomous and polytomous responses that utilize the logistic link function.

### Rasch model

The Rasch model (Rasch, 1960), sometimes referred to as the one parameter logistic model (1PL), assumes the log-odds (logit) of the item response function is a linear function of  $\theta$  and that the slopes of these linear functions are equal across all items.

$$\begin{aligned} \text{logit}\{P_j(\theta)\} &= \alpha(\theta - \beta_j) \\ P_j(\theta) &= \frac{1}{1 + \exp\{\alpha(\beta_j - \theta)\}} \end{aligned} \quad (2)$$

The intercepts ( $-\alpha\beta_j$ ) are parameterized with a negative sign so that the parameters  $\beta_j$  can be interpreted as the *difficulty* of the item; items with large values of  $\beta_j$  have lower proportions of individuals endorsing them.

The *discrimination* parameter  $\alpha$  can be fixed to some arbitrary value without affecting the likelihood as long as the scale of the individuals' propensities is allowed to be free. Common values for the discrimination are  $\alpha = 1$  and  $\alpha = 1.7$ , which is used so that the item response function is similar to the normal CDF (the standard deviation of the logistic distribution is  $\frac{\pi}{\sqrt{3}} \approx 1.8$  and a MacLauren expansion yields the approximation  $\text{logit}\{\Phi(x)\} \approx 1.6x$ ).

Three Rasch item response functions with slope  $\alpha = 1$  and difficulties  $\beta_1 = -1$ ,  $\beta_2 = 0$ ,  $\beta_3 = 1$  appear in Figure 1.

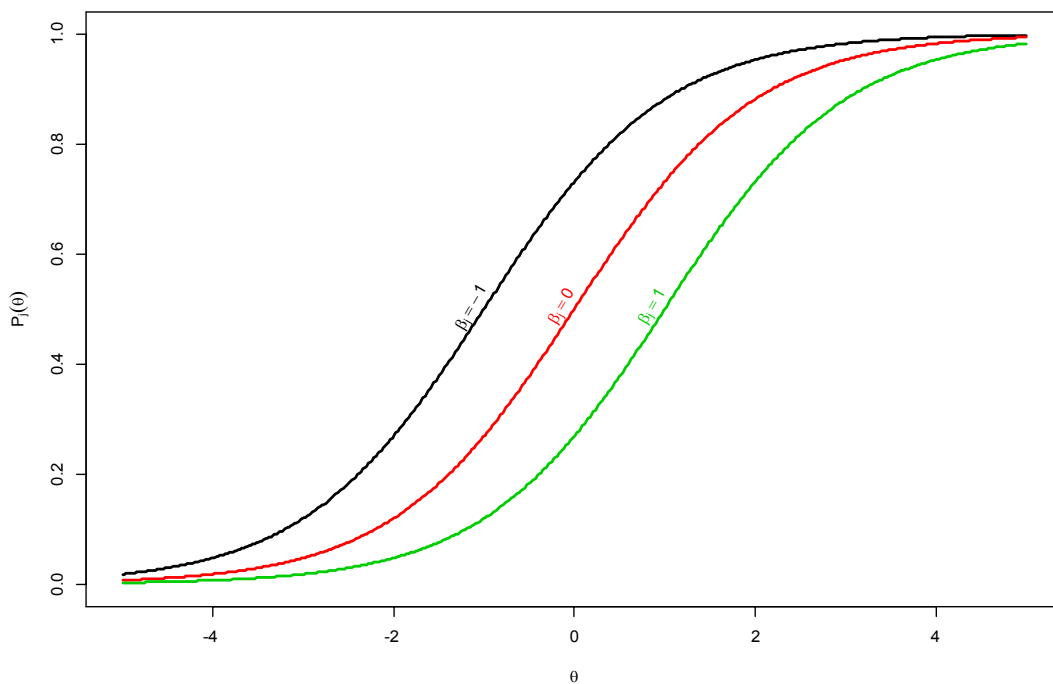


Figure 1. Three examples of the Rasch item response function with slope  $\alpha = 1$  and difficulties  $\beta = -1, 0, 1$ .

The item response functions do not intersect. This property is called the *invariant item ordering* (IIO, Sijtsma and Junker, 1996) property. The IIO property ensures that if item  $k$  is more difficult than item  $j$  (i.e.  $\beta_k > \beta_j$ ), then  $P_j(\theta) > P_k(\theta)$  for all values of the propensity  $\theta$ . Mokken (1971) refers to models like the Rasch model, that satisfy U, CI, M and IIO as double monotonicity models.

The invariant item ordering property is related to another property of the Rasch model called *specific objectivity* (Rasch, 1960; Fischer, 1987; Salzberger, 2002). Comparisons between two individuals (items) are independent of the items (individuals) used to measure them. Similarly the difference between the item difficulties of two items can be made independent of the individual  $i$  chosen for comparison. Proponents of the Rasch model claim that any method of measurement should be specifically objective, and that the Rasch model is the only IRT model that

has this property.

Another attractive property of the Rasch model is that the raw score  $S_i = \sum_{j=1}^J X_{ij}$  is a minimal sufficient statistic for the individual propensity parameter  $\theta_i$ . In fact, the Rasch model is the only possible item response model for which there exists a one-dimensional minimal sufficient statistic for the propensity parameter (Anderson, 1977).

The Rasch model is a relatively simple model with attractive properties. However, it does not fit all item response data sets. As with any statistical model, when the model does not fit the data, it should not be used for analysis. Johnson (2004) examines whether or not the Rasch model adequately explains the response behaviors of the sampled respondents to the current food insecurity items and determines that it does not.

There are two schools of thought as to how to proceed when the Rasch model does not fit a given data set. Strong believers in the Rasch model and specific objectivity argue that the Rasch model is the only model that should be used for measurement, and suggest that items that do not fit the model be discarded. Statistical modelers on the other hand attempt to expand the model so that it fits the data.

### Two parameter logistic model

In many situations the assumption that item discriminations are constant across items is too restrictive. Birnbaum (1968) introduces a model called the two-parameter logistic (2PL) model that generalizes the Rasch model by allowing the slopes to vary. Specifically the 2PL assumes the following

$$\begin{aligned} \text{logit}\{P_j(\theta)\} &= \alpha_j(\theta - \beta_j) \\ P_j(\theta) &= \frac{1}{1 + \exp\{\alpha_j(\beta_j - \theta)\}} \end{aligned} \quad (3)$$

The slope parameter, sometimes called the discrimination of the item, is a measure of how much information an item provides about the latent variable  $\theta$ . As  $\alpha \rightarrow \infty$  the item response function approaches a step function with a jump at  $\beta_j$ ; such item response functions are sometimes referred to as Guttman items (Guttman, 1950). Three 2PL items with slopes  $\alpha = 0.2, 1, 2$  and difficulties  $\beta_1 = -1, \beta_2 = 0, \beta_3 = 1$  appear in Figure 2.

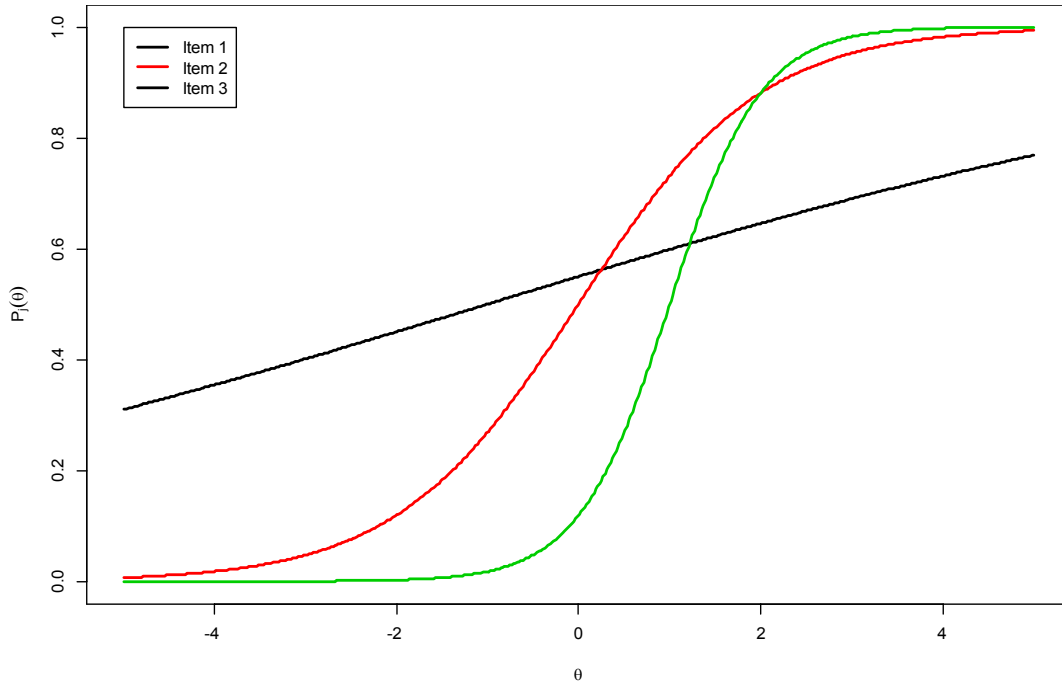


Figure 2. Three examples of the 2PL item response function with slope  $\alpha = 0.2, 1, 2$  and difficulties  $\beta = -1, 0, 1$ .

The 2PL model is not specifically objective in the sense of (Rasch, 1960). Namely, the differences between the logits of the response functions do not yield independent comparisons of individuals' propensities under the 2PL model.

Irtel (1994, 1995) extends Rasch's concept of specific objectivity to the two-parameter logistic model when discrimination parameters are unknown. Irtel's extension does not allow direct comparisons to be made between two individuals, rather it allows for comparisons between the individuals with respect to a third reference respondent.

The 2PL does not have a simple sufficient statistic for the propensity parameters, unless the discrimination parameters are fixed and known. When the discrimination parameters  $\alpha_j$  are known the weighted raw score  $S_i^{\hat{\alpha}} = \sum_j \alpha_j X_{ij}$  is a minimal sufficient statistic for the propensity  $\theta_i$ . However, the discrimination parameters are rarely known in advance and must be estimated.

Johnson (2004) presents evidence suggesting that the Rasch model may not be adequate for the analysis of the food insecurity items by demonstrating that the item slopes estimated for the 2PL were significantly different for some pairs of items.

### Other IRT models for dichotomous data

**Three parameter logistic model.** The response functions  $P_j(\theta) \rightarrow 1$  as  $\theta \rightarrow \infty$  and  $P_j(\theta) \rightarrow 0$  as  $\theta \rightarrow -\infty$  for both the Rasch and 2PL models. However, for multiple choice test items, cognitive theory suggests that when an examinee does not know the correct response, the individual will guess. In situations where guessing is possible, the assumption  $\lim_{\theta \rightarrow -\infty} P_j(\theta) = 0$  is not a

reasonable assumption of the cognitive process the model is attempting to measure. For this reason Birnbaum (1968) developed a generalization of the 2PL that allows the IRF  $P_j(\theta)$  to have a lower asymptote different from zero. The generalization is

$$P_j(\theta) = \gamma_j + \frac{1 - \gamma_j}{1 + \exp\{\alpha_j(\beta_j - \theta)\}} \quad (4)$$

The 3PL assumes that the examinee knows the correct answer of the item with probability equal to (3) or guesses the item correctly with probability  $\gamma_j$ . Figure 3 contains three 3PL item response functions with slopes  $\alpha = 0.2, 1, 2$ , difficulties  $\beta = -1, 0, 1$  and asymptotes  $\gamma = 0.4, 0.2, 0.3$ . As desired, these IRFs do not approach zero as the propensity  $\theta$  approaches  $-\infty$ .

The 3PL model may be useful in applications other than educational testing. In many attitudinal surveys, there are items for which it makes sense to assume that all individuals have a probability that is bounded below by some non-zero number  $\gamma$ , regardless of the individual's propensity.

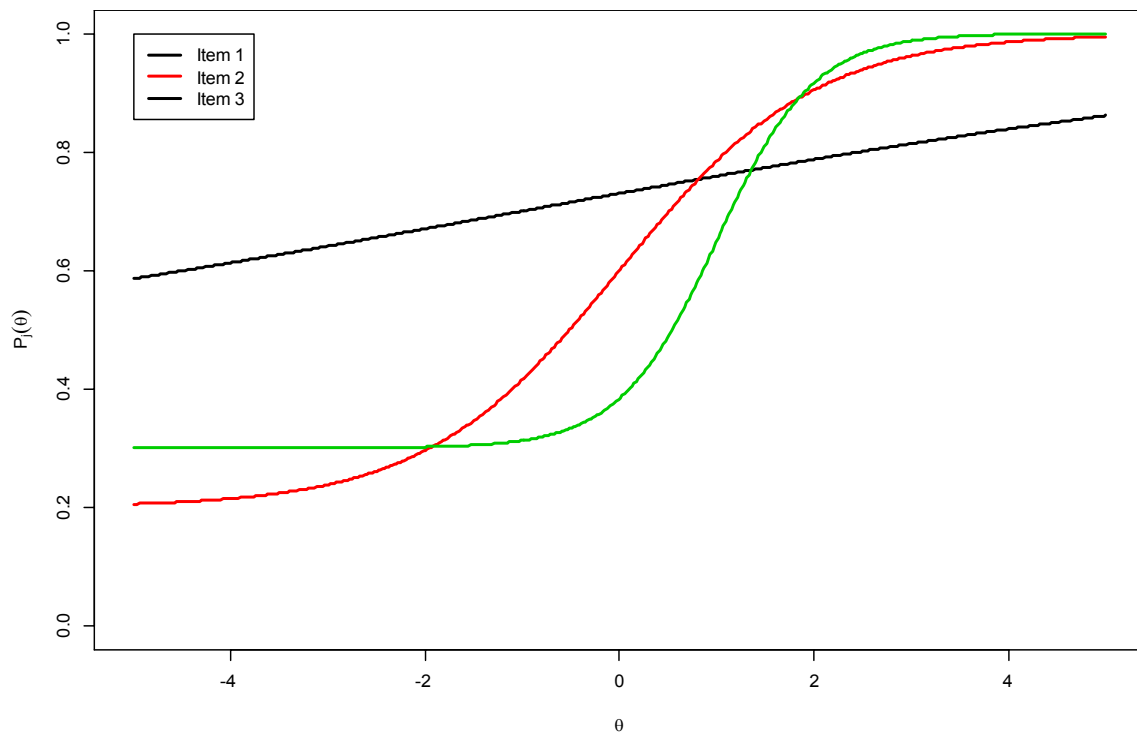


Figure 3. Three examples of the 3PL item response function with slopes  $\alpha = 0.2, 1, 2$ , difficulties  $\beta = -1, 0, 1$  and asymptotes  $\gamma = 0.4, 0.2, 0.3$ .

**Non-parametric IRT models.** Many researchers have suggested using the total score  $S_i$  as the

independent variables in a non-parametric logistic regression as a way to examine the shape of the unknown response function  $P_j(\theta)$ . Ramsay (1991), for example, uses Kernel regression as a way to estimate  $P_j(\theta)$ . Although Douglas (1997) shows that this method consistently estimates both the shape of the item response function and the rank order of examinees, the method does not work well for small data sets.

Ramsay and Abrahamowicz (1989) and Winsberg, Thissen, and Wainer (1984) on the other hand suggest methods for the estimation of non-parametric response functions  $P_j(\theta)$ , which utilize B-splines. The B-spline item response model is likely too complicated to use operationally. However, it can be utilized to examine the appropriateness of the simpler 1-, 2-, and 3-parameter item response models. Three B-spline response functions generated by assuming the logit of  $P_j(\theta)$  is a monotone B-spline function are displayed in Figure 4.

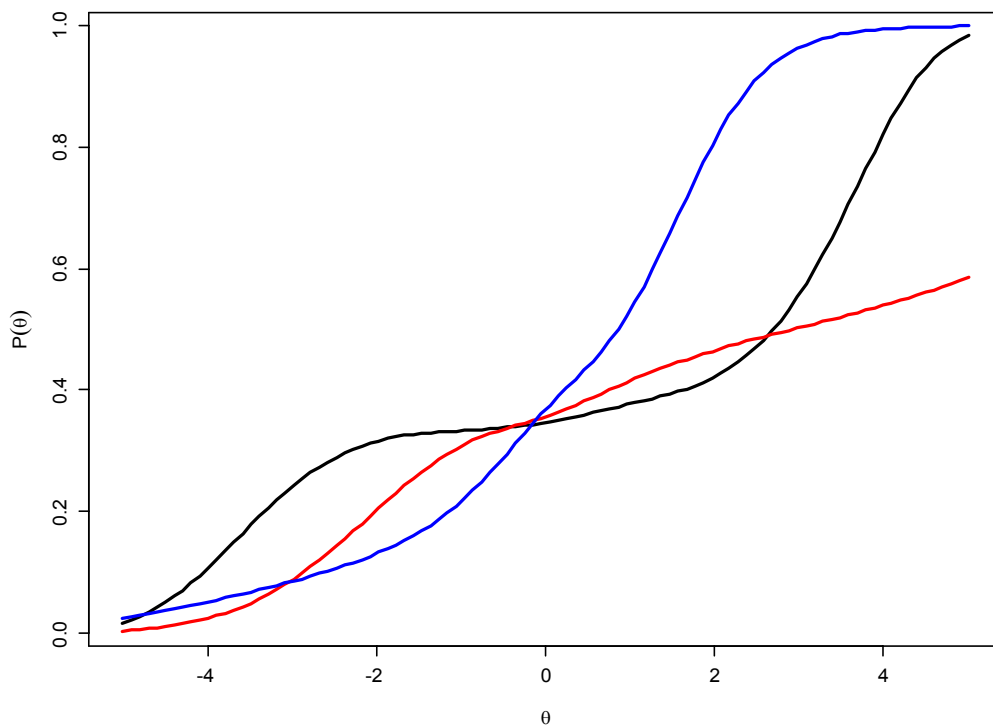


Figure 4. Three item response functions generate from a B-spline function with knots located at -1, 0, and 1.

Jim Ramsay provides free software for fitting nonparametric IRT models. The software is available at [www.psych.mcgill.ca/faculty/ramsay/TestGraf.html](http://www.psych.mcgill.ca/faculty/ramsay/TestGraf.html).

### Item response models for polytomous data

A number of questions on the food security survey are scored on a polytomous scale. However, in analysis the polytomous responses are collapsed to form dichotomous items. Although the collapsing of categories does not violate any of the core assumptions of IRT (unidimensionality, monotonicity, conditional independence), it does throw away information that could prove valuable for the classification of individuals as food insecure and/or food insecure with hunger.

Figure 5 displays the information curves for a trichotomous item both before (black curve) and after collapsing the upper two categories.

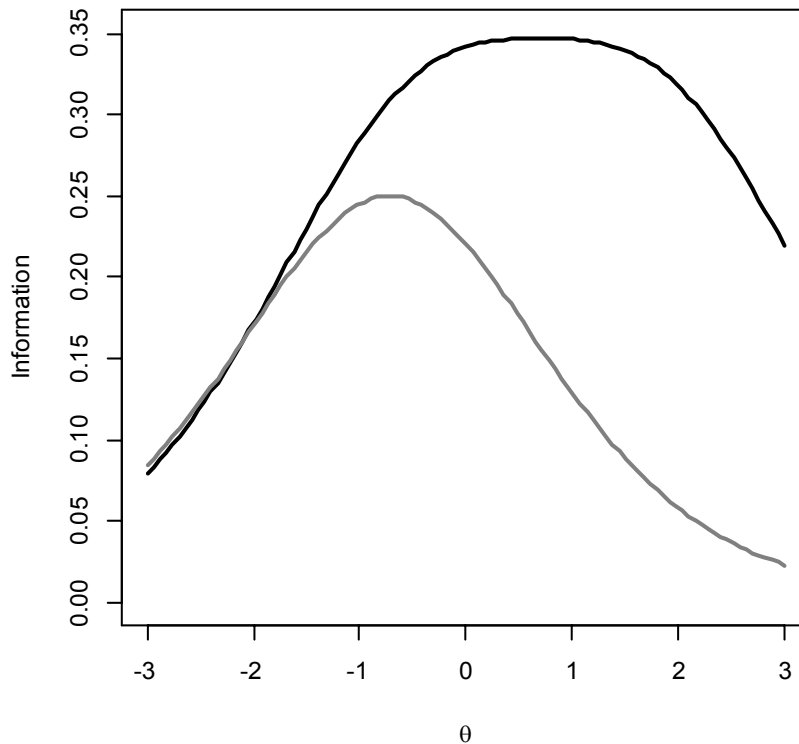


Figure 5. The information about the latent proficiency from a trichotomous partial credit item response before (black curve) and after (gray curve) collapsing the top two categories.

The parameters used to create these information curves are similar to what might be expected if a partial credit model was fit to the item on the food security survey that asks the respondent if he or she was ever worried food would run out. It is clear from the figure that the polytomous item is far superior in the amount of information it provides about the underlying propensity  $\theta$ .

It may be advantageous to use a response model for polytomous data when analyzing the food security survey data. Several authors have suggested item response models for such items. In the sections below I review three such response models. Each response model can be viewed as a generalization of the 2PL and/or Rasch model for polytomous data; all three assume some function of the item-category response functions  $P_{jm}(\theta) = Pr[X_{ij} = m | \theta_i]$  is linear in the propensity  $\theta$ . The descriptions assume that item  $j$  is scored with the  $k_j + 1$  integers from 0 to  $k_j$ .

### Graded response model

The graded response model (GRM; Samejima, 1969) assumes that the log-odds of scoring  $m$  or higher on item  $j$  is a linear function of the latent propensity  $\theta$

$$\log \left\{ \frac{Pr[X_{ij} \geq m | \theta]}{Pr[X_{ij} < m | \theta]} \right\} = \alpha_j (\theta - \beta_{jm}).$$

Unlike the response models discussed below, the graded response model requires that the discrimination parameters are fixed across item categories and that the item-category step parameters  $\beta_{jm}$  are ordered by the category index  $m$ ,  $\beta_{j1} < \beta_{j2} < \dots < \beta_{jk_j}$ .

**Partial credit model.** *The partial credit model (PCM; Masters, 1982; Muraki, 1992) assumes that the adjacent category logits are a linear function of the propensity  $\theta$ :*

$$\log \left\{ \frac{Pr[X_{ij} = m | \theta, X_{ij} \in \{m, m-1\}]}{Pr[X_{ij} = m-1 | \theta, X_{ij} \in \{m, m-1\}]} \right\} = \alpha_{jm} (\theta - \beta_{jm}),$$

which leads to the following item-category response functions:

$$\begin{aligned} P_{jm}(\theta_i) &= Pr\{X_{ij} = m | \theta_i\} \\ &= \frac{\exp\{\sum_{\ell=0}^m \alpha_{j\ell} (\theta_i - \eta_{j\ell})\}}{\sum_{r=0}^{k_j} \exp\{\sum_{\ell=0}^r \alpha_{j\ell} (\theta_i - \beta_{j\ell})\}} \end{aligned}$$

Typically researchers assume that the item-category discrimination parameters are constant across categories (i.e.  $\alpha_{jm} = \alpha_j$ ). When the discriminations are allowed to vary across items, the resulting model is referred to as the generalized partial credit model (GPCM).

Opsomer, Jensen, Nusser, Dirgei, and Amemiya (2002) fit the partial credit model (with constant discrimination parameters  $\alpha_{jm} = \alpha$ ) to data from the food security data, but found that the model did not fit the data very well. It may be possible to overcome the problems with the fit of the model by allowing the discrimination parameters to vary across items.

**Rating scale model.** *The rating scale model (RSM) assumes that the continuation logits are linear in the propensity score:*

$$\log \left\{ \frac{Pr[X_{ij} \geq m | \theta, X_{ij} \geq m-1]}{Pr[X_{ij} = m-1 | \theta, X_{ij} \geq m-1]} \right\} = \alpha_{jm} (\theta - \beta_{jm}),$$

Assuming that the continuation logits are linear results in the following item-category response functions:

$$P_{jm}(\theta_i) = \frac{\exp\{\sum_{\ell=1}^m \alpha_{j\ell} (\theta_i - \beta_{j\ell})\}}{\prod_{\ell=1}^m (1 + \exp\{\alpha_{j\ell} (\theta_i - \beta_{j\ell})\})}$$

One interesting property of the rating scale model is that for each  $(k_j + 1)$ -category item,  $k_j$

dichotomous pseudo-items can be created that are conditionally independent under the model. The  $m$  th pseudo item for respondent  $i$ , denoted  $X_{ij_m}$  is coded as follows:

$$X_{ij_m} = \begin{cases} \text{missing} & \text{if } X_{ij} < m - 1 \\ 0 & \text{if } X_{ij} = m - 1 \\ 1 & \text{if } X_{ij} \geq m \end{cases}$$

These pseudo-items can then be analyzed using one of the previously mentioned IRT models for dichotomous data.

In most applications all three of the models discussed above for polytomous items will fit the data equally well (as long as all three are using the same number of parameters), and so the choice between the three comes down to the personal choice of the researcher.

### Modeling the population distribution of propensities

The mixed effects versions of IRT models are defined by the item response functions and the propensity distribution  $F(\theta)$ . Most applications utilizing measurement scales are interested in the latent propensities ( $\theta$ ) of the individuals in the sample, e.g., ranking examinees according to their propensities. The Food Insecurity program is more interested in how those propensities are distributed across the population, and therefore, is interested in the distribution  $F(\theta)$ . Typically IRT modelers assume that the distribution  $F$  is the normal distribution with mean zero and standard deviation one. However, the normal distribution does not necessarily work for all applications and care must be taken when choosing the parametric form of  $F$ .

In the 2002 Food Security survey approximately 90% of the sampled respondents were removed from the study because their responses to preliminary questions indicated that they would most likely not be food insecure (e.g., their income level was too high). So, if experts agree that the *entire* population of propensities is normally distributed, then the mixing distribution used in analysis must account for the non-representative sample. One possible correction, implemented in Johnson (2004) uses a normal distribution truncated at the 90th percentile (approximately 1.28):

$$F(\theta) = \begin{cases} \frac{\Phi(\theta) - \Phi(1.28)}{1 - \Phi(1.28)} & \text{if } \theta > 1.28 \\ 0 & \text{otherwise} \end{cases}$$

A similar approach was employed by Nord (1999; as cited in Opsomer, Jensen, Nusser, Dringei, and Amemiya, 2002) to estimate the distribution of food insecurity in all U.S. households.

Another way to get around the difficulty of defining a mixing distribution is to assume some non- or semi-parametric form. For example, the analysis of the National Assessment of Educational Progress, a large scale educational survey, assumes examinee propensities  $\theta_i$  are independently and identically distributed according to a discrete distribution on 41 equally spaced points from  $-4$  to  $4$  with unknown mass. That is, the probability mass function for the propen-

sity  $\theta$  is

$$f_{\Theta}(t) = \begin{cases} p_t & \text{if } t \in \{-4, -3.8, \dots, 4\} \\ 0 & \text{otherwise} \end{cases}$$

where  $\sum_i p_i = 1$ , and the mean and variance of the distribution are constrained to be zero and one respectively. Mislevy and Bock (1982) and Muraki and Bock (1997) provide more information on the use of this discrete distribution in the analysis of item response data.

Mixture models have the added flexibility that other information about the respondents can be incorporated into the mixing distribution. Suppose for example, that researchers would like to determine how food security varies by income, age, and race, and that this information is contained in the independent variable  $\mathbf{y}_i$ . If the propensities were known then a logical step in analysis would be to perform a linear regression of  $\theta$  on  $\mathbf{y}_i$ .

Although the propensities are latent, this regression can still be performed by assuming the mean of the distribution for individual  $i$  is a linear function of the independent variable  $\mathbf{y}_i$ . In the normal case, this would amount to  $\theta_i \sim N(\mathbf{y}_i^t \boldsymbol{\gamma}, 1)$ . The estimates of  $\boldsymbol{\gamma}$  from this latent regression can be interpreted directly without having to rely on the propensities of all individuals in the sample, and the population distribution for the entire population is the mixture distribution

$$F(\theta) = \sum_{\mathbf{y}} F(\theta | \mathbf{y}) \pi(\mathbf{y}),$$

where  $\pi(\mathbf{y})$  is the proportion of the population with background characteristics defined by the vector  $\mathbf{y}$ . Opsomer, Jensen, and Pan (2002) perform a latent regression analysis of food security data.

In the sections that follow I will assume that the distribution is defined by some set of parameters denote  $\eta$ . The parameters could be the mean and variance of a normal distribution, the regression effects associated with the latent regression, or the masses  $p_i$  in the discrete distribution discussed above. I will denote the cumulative distribution function by  $F_{\eta}$  and the density by  $f_{\eta}$  to remind the reader.

### Estimating univariate IRT models

Johnson (2004) reviews the four basic techniques for the estimation of item response models: joint maximum likelihood, conditional maximum likelihood, marginal maximum likelihood, and Bayesian estimation with Markov chain Monte Carlo. All four methods rely heavily on the assumption that individuals are independent of one another, and that the item responses of a given individual are independent given that individual's propensity score  $\theta_i$ . Under the assumption of conditional independence the joint probability of the item response vector  $\mathbf{x}_i$  conditional on  $\theta_i$  is

$$L_i(\theta | \mathbf{x}_i, \boldsymbol{\psi}) = Pr\{\mathbf{x}_i | \theta_i, \boldsymbol{\psi}\} = \prod_{j=1}^J Pr\{X_{ij} = x_{ij} | \theta_i, \boldsymbol{\psi}_j\}, \quad (5)$$

where  $\boldsymbol{\psi}_j$  is the vector of all item parameters for item  $j$ . For example, the likelihood for propensity  $\theta$  under the 2PL model, where  $\boldsymbol{\psi}_j = (\alpha_j, \beta_j)^t$ , is:

$$L_i(\theta_i | \mathbf{x}_i, \boldsymbol{\psi}) = \frac{\exp\{\theta_i \sum_j x_{ij} \alpha_j - \sum_j x_{ij} \alpha_j \beta_j\}}{\prod_j [1 + \exp\{\alpha_j (\theta_i - \beta_j)\}]}$$

The following summarizes the comments in Johnson (2004) about each of the estimation procedures.

- The conditional maximum likelihood procedure is only applicable for the simplest IRT models, namely the Rasch model for dichotomous data and the partial credit model for polytomous data.
- The JML procedure estimates the item parameters ( $\boldsymbol{\psi}$ ) and examinee abilities by maximizing  $L(\boldsymbol{\psi}, \theta; \mathbf{X}) = \prod_i L_i(\theta | \mathbf{x}_i, \boldsymbol{\psi})$  with respect to  $\boldsymbol{\psi}$  and  $\theta$  simultaneously. One of the problems with JML estimates in models similar to IRT models is that the estimates are inconsistent (Neyman and Scott, 1948; Andersen, 1970; Ghosh, 1995). In terms of IRT models, this means that no matter how many individuals are included in the sample, the estimates for the item parameters may still be biased.
- Integrating the random effects (i.e. propensities) out of the individual likelihoods defined in (5) defines the marginal probability of observing the item response vector  $\mathbf{x}_i$ ,

$$Pr\{\mathbf{x}_i | \boldsymbol{\psi}, \boldsymbol{\eta}\} = \int_{\Theta} L_i(\theta | \mathbf{x}_i, \boldsymbol{\psi}) dF_{\boldsymbol{\eta}}(\theta). \quad (6)$$

- Taking the product of the probabilities in (6) over individuals  $i$  defines the marginal likelihood of the parameters  $\boldsymbol{\psi}$  and  $\boldsymbol{\eta}$

$$L(\boldsymbol{\psi}, \boldsymbol{\eta} | \mathbf{X}) = \prod_i Pr\{\mathbf{x}_i | \boldsymbol{\psi}\}, \quad (7)$$

- which is maximized with respect to the item parameters  $\boldsymbol{\psi}$  and population parameter  $\boldsymbol{\eta}$  to derive the MML estimates.
- Once the MML estimates  $\hat{\boldsymbol{\psi}}$  and  $\hat{\boldsymbol{\eta}}$  have been obtained, the posterior distributions of each individual's propensity is approximated using an empirical Bayes procedure. The posterior density function for respondent  $i$ 's propensity score is approximated by calculating

$$f(\theta_i | \mathbf{x}_i) \propto L_i(\theta_i | \mathbf{x}_i, \hat{\boldsymbol{\psi}}) f_{\hat{\boldsymbol{\eta}}}(\theta_i)$$

- where the proportionality constant forces the posterior density  $f(\theta_i | \mathbf{x}_i)$  to integrate to one. If a point estimate is desired for each respondent we can approximate the mean, median or mode of the empirical Bayes approximated posterior distribution and use the approximation as the point estimate of the respondent's propensity score.
- The Bayesian method for estimation of IRT models is similar to the marginal likelihood technique described in the previous section. However, in addition to assuming a mixing distribution for the propensities, Bayesian analysis places a prior distribution on each of the model parameters. It is also possible to simultaneously estimate posterior quantities for both the items and the respondents in the data set, so there is no need for an empirical Bayes procedure like the one described above.

One of the shortcomings of a Bayesian analysis of an IRT model is that numerical integration techniques must be used to approximate the posterior distributions (Patz and Junker, 1999). The numerical method, called Markov chain Monte Carlo (MCMC), can be quite time consuming for large data sets, and requires extreme care to make sure that the resulting estimates are valid.

### Multidimensional IRT models

As noted earlier food insecurity data has suggested that items asking specifically about children appear to be measuring a different, but related construct, when compared to items asking about the adults in a household. There are essentially three ways to handle the problem of multidimensionality: (a) ignore the problem and proceed with a univariate analysis; (b) discard items loading on the dimension deemed least important; or (c) model the multidimensionality. Ignoring the problem makes the resulting scale difficult to interpret, and discarding items is throwing away useful information, so solution (c) is probably preferable.

The simplest model for multidimensional item response data is the simple structure model, which assumes that each item is affected by only a single dimension. For example, for the food security data we assume that there is one propensity  $\nu_{i1}$  associated with adult items, and a second propensity  $\nu_{i2}$  associated with the items asking about the children in a household. The individual-level likelihood for the propensity in (5) becomes the following function of  $\nu_{i1}$  and  $\nu_{i2}$

$$L_i(\nu_{i1}, \nu_{i2} | \mathbf{x}_i, \psi) = \prod_{j \in G_1} Pr\{X_{ij} = 1 | \psi_j, \nu_{i1}\} \prod_{j \in G_2} Pr\{X_{ij} = 1 | \psi_j, \nu_{i2}\},$$

where  $G_1$  and  $G_2$  are sets containing the indices of items loading on  $\nu_{i1}$  and  $\nu_{i2}$  respectively. The joint maximum likelihood estimates can then be found by maximizing  $\prod_i L_i(\nu_{i1}, \nu_{i2} | \mathbf{x}_i, \psi)$  with respect to the  $\nu$ 's and the item parameter in  $\psi$  simultaneously, which is equivalent to performing two univariate JML estimations, one for each  $\nu$ .

For marginal maximum likelihood and Bayesian estimation, the population model  $F_\eta$  is a bivariate distribution. The individual likelihoods in (6) require the evaluation of the double integral

$$Pr\{\mathbf{x}_i | \psi, \eta\} = \int_{\nu_1} \int_{\nu_2} L_i(\nu_1, \nu_2 | \mathbf{x}_i, \psi) dF_\eta(\nu_1, \nu_2).$$

The approximation of double integrals is quite a bit more computationally intensive than single integrals. Normally software packages for estimating univariate IRT models approximate integrals using anywhere from 10 to 50 quadrature points. If the same number of points is used for each dimension in a bivariate analysis the individual likelihoods would have to be evaluated for each of 100 to 2500 points in the plane. If there are 10,000 respondents in the sample, then each iteration of the estimation algorithm would require 1 to 25 million calculations.

Another difficulty that is encountered with multidimensional IRT models is deciding how to interpret the results. If a bivariate model was fit to the food security data we would be left the question, “Which dimension is the ‘Food Insecurity’ dimension?” Maybe we would like to create a weighted average of the  $\nu$ ’s to define our food security, but we must decide what the weights should be.

One way to choose the weights is to perform a latent factor analysis on the two propensities  $\nu_1$  and  $\nu_2$ . This could either be done explicitly in the definition of the bivariate distribution  $F_\eta$  (see for example the testlet model of Bradlow et al., 1999) or a *post hoc* analysis could be performed by drawing a number of *plausible values* from each respondents posterior distribution  $F(\nu_1, \nu_2 | \mathbf{x}_i)$  (or its empirical Bayes approximation) and performing factor analyses on the plausible values.

## EVALUATING THE ITEMS/MODELS

Section 3 reviewed a number of models for the analysis of item response data. It is important to determine whether the data actually resembles data that comes from a model that can be closely approximated with the model selected for analysis. Any model used to measure food insecurity and/or hunger should be evaluated before inferences about the prevalence of food insecurity are made. Section 1 discusses methods for selecting the best model from a small set of possible models, Section 2 reviews methods for diagnosing problems with a given model, and Section 3 discusses concerns with missing data.

### Model selection

One way to examine the efficacy of a given model is to compare it with an alternative that is slightly more complex. For example, before using the Rasch model for analysis, the fit of the Rasch model should be compared with a 2PL.

When the smaller model (e.g., Rasch) is nested within the larger model (e.g., 2PL) the likelihood ratio statistic can be utilized. The likelihood ratio statistic is calculated by finding the maximum attainable value of the likelihood under the two models, and then calculating the ratio of the two maximal likelihoods:

$$\lambda_{12}(\mathbf{X}) = \frac{L(\hat{\psi}_1 | \mathbf{X})}{L(\hat{\psi}_2 | \mathbf{X})},$$

where  $\hat{\psi}_k$  is the vector of maximum likelihood estimates under model  $k$ .

For large sample sizes the distribution of the statistic  $-2 \log \lambda_{12}(\mathbf{X})$  approaches a chi-squared ( $\chi^2$ ) distribution with the number of degrees of freedom equaling the difference in the number of free parameters under the two models. The smaller model would be rejected in favor of the larger model for large values of  $-2 \log_{12}(\mathbf{X})$ . The statistic for comparing the Rasch model to the 2PL would approach a  $\chi^2$  random variable with  $J - 1$  degrees of freedom, where  $J$  is the total number of items. If the  $\chi^2$  approximation was used to compare the 2PL fit to the Rasch fit for the ten household questions, the Rasch model would be rejected if  $-2 \log_{12}(\mathbf{X}) - 16.9 > 0$  (at the 0.05 significance level).

The Akaike and Bayesian (or Schwarz) information criteria (AIC & BIC) compare the fit of two models, whether the models are nested or not. Both criteria penalize models for the number of parameters included in the models. The AIC penalizes the log of the likelihood ratio statistic by subtracting twice the difference in the number of parameters

$$AIC_{12} = -2 \log \lambda_{12}(\mathbf{X}) - 2q_{12},$$

where  $q_{12}$  is the difference in the number of parameters between the models (e.g.,  $J - 1$  when comparing Rasch with 2PL). The larger model is selected if the AIC is larger than zero. For the ten household items, the AIC would choose the 2PL model over the Rasch model whenever  $-2 \log \lambda_{12}(\mathbf{X}) - 18 > 0$ , so it is less likely to choose the 2PL than the  $\chi^2$  approximation.

The Bayesian information criterion is even more conservative. The BIC also penalizes for the number of parameters in the larger model, but weights the penalty factor more heavily for large data sets. The BIC for comparing two models is defined

$$BIC = -2 \log \lambda_{12} - q_{12} \log N,$$

where  $N$  is the number of respondents in the sample. Like the AIC, the BIC selects the larger model whenever the BIC is positive. If 10,000 respondents are utilized in the analysis of the food insecurity items, then the penalty term for comparing the 2PL to the Rasch for the ten household items is approximately  $q_{12} \log 10,000 \approx 83$ .

Ideally we would want to compare the fits of all possible IRT models, i.e., all models that satisfy the unidimensionality, conditional independence, and monotonicity assumptions. The class of all such models is infinite, so it is impossible to perform such an analysis. However, if we compare to more flexible models, such as the spline-based IRT models of Ramsay and Abrahamowicz (1989) and Winsberg, Thissen, and Wainer (1984), and determine that the simple model (e.g, 2PL or Rasch) fits no worse, then it may be reasonable to proceed with the simple model.

### Diagnostics for assessing model fit

It is also possible to perform diagnostic analyses to examine the fit of various item response models to the data. These analyses compare the observed data to the fitted model using a number of statistics meant to measure the discrepancy between the assumed model and the observed data. Although a great deal of research in IRT has concentrated on assessing model fit, there is not a set of universally accepted discrepancy measures. Sinharay and Johnson (2003) recommend the following discrepancy measures for dichotomous item response models:

- *Biserial correlation.* The point biserial correlation is the correlation between an item score  $X_{ij}$  and the total score  $S_i = \sum_j X_{ij}$ ; the rest-biserial correlation, which measure the correlation between  $X_{ij}$  and the *rest-score*  $S_i - X_{ij}$ , is also a useful measure. The biserial correlation for item  $j$  is

$$r_j = \frac{r_j^{(pb)} \sqrt{\hat{p}_j(1 - \hat{p}_j)}}{\phi(\Phi^{-1}(\hat{p}_j))},$$

- where  $r_j^{(pb)}$  is the point biserial correlation for item  $j$ ,  $\hat{p}_j$  proportion of the sample that affirmed item  $j$ , and  $\phi$  is the standard normal density function, and  $\Phi$  is the standard normal cumulative distribution function.
- The biserial correlations are relatively constant across items generated from a Rasch model, and hence, the variance of the of the biserial correlations provides an overall measure of the adequacy of the Rasch model for a given data set. If the variance of the biserial correlations is small then the Rasch is probably adequate. If the variance is large another model should be explored. Sinharay and Johnson (2003) also note that the biserial correlations detect the misfit of 2PL models fit to the 3PL data. However, whether the biserial correlations are good for detecting the misfit of models like the 2PL fit to data generated from IRT with more complicated item response functions has yet to be examined.
- *Odds ratios.* The odds ratio between two dichotomous items is

$$OR_{ij} = \frac{n_{11}^{(i,j)} n_{00}^{(i,j)}}{n_{10}^{(i,j)} n_{01}^{(i,j)}},$$

- where  $n_{kk}^{(i,j)}$  is the number of respondents with a score of  $k$  on item  $i$  and a score of  $k'$  on item  $j$ .
- The odds ratios should be relatively constant between pairs of items, and hence, the variance of the odds ratios, either within an item, or across all items, is a useful measure for detecting misfit of the Rasch model. Although Sinharay and Johnson (2003) found that the odds ratios were not useful measures for detecting the violations of the shape assumptions of the 2PL, they were found to be useful for detecting violations of unidimensionality and local independence.

- To determine whether the observed data behaves differently than a specific model, several data sets must be generated from the assumed model, and the discrepancy measures are calculated for each of the simulated data sets. Two methods have been suggested for simulating data from the assumed model:
- *The posterior predictive method.* The posterior predictive method for performing diagnostic checks of the fit of a specified model to a given data set generates the data from the posterior predictive distribution. Data can be generated from the posterior predictive distribution by sampling the parameter values from their joint posterior distribution and then drawing data from the likelihood defined by these parameter values.
- For example, for simulated data set number  $t$ , the item parameters  $\psi^{(t)}$  and the distribution parameters  $\eta^{(t)}$  would be drawn from the joint posterior distribution  $\pi(\psi, \eta | \mathbf{X})$ . Then  $N$  propensities would be drawn from the distribution defined by the parameters  $\eta^{(t)}$ , i.e.,  $\theta_i^{(t)} \square F(\theta | \eta^{(t)})$  for all  $i = 1, \dots, N$ . Item responses are then generated for each individual and each item using the item parameters in  $\psi^{(t)}$ .
- For more information on the posterior predictive method for assessing the fit of IRT models see Sinharay and Johnson (2003).
- *The parametric bootstrap.* Most practitioners prefer to perform maximum likelihood estimation, and hence the posterior predictive method is not an option, because it requires the posterior distribution of the model parameters. The parametric bootstrap, rather than sampling from the predictive distribution, samples data directly from the fitted model with all parameter values set to their point estimates, usually their MLEs. It is a sort of empirical Bayes version of posterior predictive checks. As long as the sample size is large, the posterior predictive method and the parametric bootstrap will not vary substantially. However, for small sample sizes the parametric bootstrap may have a difficult time detecting misfit.

Once the data sets have been generated the discrepancies calculated from the observed data are compared to those from the simulated data sets. Comparisons are typically made by calculating the proportion of samples that have discrepancy measures larger than the one calculated from the observed data. If the observed discrepancies are more extreme than a large number (e.g., 95%) of the simulated discrepancies, then there is reason for concern; the assumed model does not adequately explain the data. The researcher then must decide whether to expand the model or attempt to remove problematic items from the analysis.

### **Assumptions about missing data**

Missing data can also be problematic when making inferences. There are two types of missing data that are present in the food security survey data. The first type of missing response is generated when an individual responds “No” to a parent item, at which time the respondent is asked to skip to a later question. The second type of missing data is produced when a respondent refuses to answer a question. In order to deal with missing data we must be able to either, (a) assume the data is missing at random; or (b) be able to model the mechanism by which data is missing.

As Opsomer, Jensen, Nusser, and Amemiya (2002) note, questions with follow-ups often violate the assumption of conditional independence. In a 2PL analysis, violations of this conditional independence assumption can artificially inflate discrimination parameters for the clustered items and deflate the discrimination parameters for the other questions in the survey. The analysis would then produce propensity scores that place inflated weights on the clustered items.

As long as the follow-up questions are treated as missing for individuals who answer “No” to the parent questions, the conditional independence assumption is probably OK. Missingness can be ignored whenever it is missing at random, that is whenever the missingness does not depend on the response that would have been given, or when the missingness mechanism can be modeled. In the case of the food security items there is no response that would make sense for individuals who respond “No” to the parent item. We know exactly the mechanism by which these individuals have received a missing response to the item:

$$Pr\{X_5 = \text{missing} \mid X_4 = 0\} = 1.$$

Hence, results will not be biased by treating the missing data as missing at random. In fact, as Opsomer, Jensen, Nusser, and Amemiya (2002) note, the scale does not change when the follow-up questions are excluded from the analysis.

Although the missingness can be modeled, interpretation of the model parameters for such items must be made carefully. The item parameters are actually for a conditional item that is only administered if a response of “Yes” was given for the previous question.

A more likely problem occurs for those individuals who refuse to respond to items. It is quite likely that individuals who are food insecure are less likely to respond to questions about food security, because they might be embarrassed or fear governmental intervention. If this is the case, then the data is not missing at random, and treating the data as such will necessarily bias the results.

### **Comparing response behaviors of households with and without children**

Froelich (2002) noted that there appeared to be two latent constructs behind the item responses for households with children. One of the constructs corresponds to the individual/household items and the other dimension corresponds to the eight items regarding children. Another important question to ask is whether households with children respond to the household questions in the same way as households without children.

One way to compare the types of households is with a differential item functioning analysis on the ten household items. If the household items are found to function differently across the two household types, then their responses on those questions are not comparable.

A second method would be to perform a model selection. One model would perform separate analyses of the two household types, and calculate the maximal likelihood function for the two sets. The reduced model would force the item parameters for the ten household items to be the same across the two household types. An approximate  $\chi^2$  test, the AIC, or the BIC would then be used to select the best fitting model. If the full model that allows item parameters to be differ-

ent across the two household types is selected, the responses to the ten household items are not comparable across the two types of household.

### PREVALENCE ESTIMATES

In order to estimate the prevalence of food insecurity and/or hunger we must be able to classify each respondent into one of the food security classes, or be able to calculate the probabilities indicating how likely each respondent is to be in each of the categories. To calculate these probabilities (or classification rules) I will assume that a unidimensional IRT model has been fit to the data, and that as part of that IRT model a population distribution  $F_{\eta}(\theta)$  was estimated.

#### Defining cutpoints on the measurement scale

Let  $\zeta_i = 0$ ,  $\zeta_i = 1$ , and  $\zeta_i = 2$  denote the realizations that individual  $i$  is food secure, food insecure, or food insecure with hunger respectively. Ideally, there would be a perfect monotone relationship between true food-security/hunger variable  $\zeta_i$  and the propensity  $\theta_i$  measured by the survey questions. If there was a perfect monotone relationship between the two, then there exists cutpoints  $\tau_1$  and  $\tau_2$  such that:

$$\zeta_i = \begin{cases} 0 & \text{if } \theta \leq \tau_1 \\ 1 & \text{if } \tau_1 < \theta \leq \tau_2 \\ 2 & \text{if } \theta > \tau_2 \end{cases} \quad (8)$$

If (8) holds, then the cutpoints could be used classify individuals in the food security survey sample and to aid in the estimation of the percentage of the entire populations in the different categories.

The relationship between  $\zeta_i$  and  $\theta_i$  could also be treated as probabilistic, where the distribution of the propensity  $\theta_i$  depends on the food security and hunger status  $\zeta_i$  of individual  $i$ . Let  $F(\theta | \zeta = z)$  denote the distribution of the propensity  $\theta$  given the individuals' food security status is  $\zeta = z$ . Hopefully the distributions are *stochastically ordered* so that

$$F(t | \zeta = 0) \geq F(t | \zeta = 1) \geq F(t | \zeta = 2) \quad (9)$$

for all values  $t$ . If the propensities are normally distributed within each of the food security groups with means  $\mu_0$ ,  $\mu_1$ , and  $\mu_2$  and standard deviations  $\sigma_0$ ,  $\sigma_1$ , and  $\sigma_2$ , where  $\mu_z = E[\theta | \zeta = z]$  and  $\sigma_z^2 = V(\theta | \zeta = z)$ , then the stochastic ordering property in (9) is satisfied if and only if  $\mu_0 \leq \mu_1 \leq \mu_2$  and  $\sigma_0 = \sigma_1 = \sigma_2$ .

Given a validity sample, where both the hunger status variable  $\zeta_i$  and the food security item responses had been observed, an analysis could be performed the relationship between the latent propensities and the food security class defined by  $\zeta$ . However, a validity study is likely to be

too costly or impossible to obtain and another method must be used to define the relationship between the latent propensities and the food security classes. In the discussion that follows I will assume that the relationship can be described with cutpoints as in (8).

In the sections that follow I will offer some ideas about how the cutpoints can be set on the food insecurity measurement scale. Section 1 attempts to map the current methodology using raw scores onto the measurement scale, Section 2 describes a method similar to one used to set proficiency levels in the National Assessment of Educational Progress, and Section 3 discusses other possibilities.

### **Mapping the current observed-score cutpoints onto a measurement scale**

The current methodology classifies all households with a total total score  $S \leq 2$  as food secure, so it seems reasonable to assume that the experts who developed the cutoff believe that there more than half of the respondents with raw scores  $S = 2$  are from food secure households, that is,  $Pr\{\theta \leq \tau_1 \mid S = 2\} \geq 0.5$ . Similarly, we should expect that the probability that a household is food insecure, given that the respondent affirmed three items on the household measurement scale should also be greater than one-half,  $Pr\{\theta \geq \tau_{12} \mid S = 3\} \geq 0.5$ . The propensity score cutpoint between the food secure and food insecure should be a number between the conditional medians of  $\theta$  given  $S = 2$  and  $S = 3$ , suggesting the following candidate for the first cutpoint

$$\tau_1 = \frac{1}{2}(\text{median}(\theta \mid S = 2) + \text{median}(\theta \mid S = 3)).$$

A logical alternative to the above suggestion defines the cutpoint using the conditional means instead of the conditional medians. Numerical analysis techniques would need to be utilized to approximate both of the suggestions above.

### **Using a NAEP-like approach to define the cutpoints**

One of the goals of the National Assessment of Educational Progress is to report the proportion of the population in each of three achievement levels (basic, proficient, advanced). The achievement levels are conceptually similar to the idea of classifying individuals into the various food security classes. Appendix I of the 1998 NAEP technical report provides a thorough summary of the procedure used to set the achievement level cutscores (U.S. Department of Education, Office of Educational Research and Improvement, 2001). The following provides a brief summary of how the method can be used to set cutpoints on the food insecurity scale.

The first step is to define the characteristics that define a “marginal” respondent in each of the classifications, that is, the group would define what characteristics separate the most secure food insecure respondent from the least secure food secure respondent. The descriptions need not be made in terms of the questions on the measurement scale, they are simply meant to be conceptual definitions, that are understandable by the experts in the standard setting.

Judges would then be asked about the responses they would expect to see from respondents who are extreme cases in each of the food insecurity levels. For dichotomously scored items the

judges would be asked to provide estimates of the proportion of “marginal” individuals that would affirm each item.

Suppose we were setting the cutpoint that separates food secure individuals from food insecure individuals. The judges would be asked to estimate the proportion of the most secure food insecure respondents that would affirm each item. Let  $r_{jk}$  denote judge  $j$ 's estimated proportion for item  $k$ .  $J \times M$  approximate cutpoints are obtained by calculating  $t_{jk} = \hat{p}_k^{-1}(r_{jk})$  for each judge  $j = 1, \dots, M$  and item  $k = 1, \dots, J$ , where  $\hat{p}_k$  is the estimated item response function for item  $k$ . The average of all  $J \times M$  approximate cutpoints could be used as the cutpoint, i.e.,  $\tau_1 = \frac{1}{JM} \sum_{j,k} t_{jk}$ ; the cutpoint could also be defined by the median of the  $t_{jk}$ 's

The approximated cutpoints  $t_{jk}$  will likely vary by item  $k$  and judge  $j$ . Let  $v_j$  denote the variance of judge  $j$ 's approximate cutpoints over items, and  $u_k$  denote the variance of item  $k$ 's approximate cutpoints over judges. It might be wise to down-weight the contribution of judges and/or items with high variability. The weighted average below

$$\tau_1 = \frac{\sum_{j,k} w_{jk} t_{jk}}{\sum_{j,k} w_{jk}},$$

where  $w_{jk} = (v_j u_k)^{-1}$ , achieves this goal. A weighted median could also be calculated and used for the cutpoint.

A similar process would be utilized to define the cutpoint separating individuals who are food insecure without hunger from those who suffer from hunger, or any other food insecurity class.

### Other methods for setting the cutpoints

The NAEP-like cutpoint setting procedure asks judges to provide expert opinion about how marginal respondents in each of the food security classes will respond to each question. A simple alternative to the NAEP procedure is to ask the judges to provide an estimate of the average response of *all* households in the class, not just the marginal households. The judges would approximate

$$\mu_{k\ell} = E[X_k | \zeta = \ell].$$

For example, to define the first cutpoint, judges would be asked to approximate the average response to item  $j$  for individuals who are food secure:

$$\mu_{k0} = E[X_k | \zeta = 0] = \frac{\int_{-\infty}^{\tau_1} \sum_{c=1}^{K_j} kP_{kc}(\theta) dF_\eta(\theta)}{\int_{-\infty}^{\tau_1} dF_\eta(\theta)},$$

which is a non-decreasing function of the cutoff  $\tau_1$ ; I will write  $\mu_{k0}(\tau_1)$  to remind the reader

that it is a function of the cutpoint. Each judge would provide an estimate of  $\mu_{k0}$  for several items indexed by  $k$ . Let judge  $j$ 's estimate for item  $k$  be denoted  $m_{kj}$ . The cutpoint for item  $k$  could then be set at the average value of  $t_{jk} = \mu_{k0}^{-1}(m_{kj})$  over all items and all judges, or by first averaging the estimated average scores for each item (denoted  $\bar{m}_k$ ) and then taking the average of their mappings onto the food insecurity scale, i.e.,

$$\tau_1 = \frac{1}{J} \sum_{k=1}^J \mu_{k0}^{-1}(\bar{m}_k)$$

Weighted averages, like those discussed in the previous section, could also be utilized to set the cutpoints.

The procedures can be generalized further by asking the judges to estimate the average value of any element-wise non-decreasing function  $g(\mathbf{x})$  of the item response vector  $\mathbf{x}$ . The response on a single item, and the total score  $s_i = \sum_k x_{ik}$  are specific cases.

Whatever procedure is utilized to produce the cutpoints needs to be closely scrutinized. The process of setting the cutpoints is probably the most subjective piece in the measurement of food insecurity and will likely have the greatest affect on the estimated proportions in each of the food security levels.

### Using cutpoints to estimate prevalence

Suppose that a one of the procedures discussed in the previous section was utilized to find the cutoffs  $\tau_1$  and  $\tau_2$  in (8). The percentage of the population in each of the three food security classes can be estimated in one of several ways. I review three methods previously discussed in Johnson (2004) here.

Similar to the current procedure each respondent's propensity  $\theta_i$  is estimated and the weighted proportion of estimated propensities in each of the three classes is calculated:

$$\square \Pr\{\text{food secure}\} = \frac{\sum_{i=1}^n w_i I\{\tilde{\theta}_i \leq \tau_1\}}{\sum_{i=1}^n w_i},$$

where  $I\{\tilde{\theta}_i\}$  is the function indicating whether or not the estimated propensity for individual  $i$  is below the first cutpoint, and  $\tilde{\theta}_i$  is the estimated propensity for respondent  $i$  (e.g., the approximated posterior mean for the respondent).

One of the drawbacks of this approach is that it treats all individuals equally, regardless of how much information we have about them. Ideally we would want to weigh individuals with complete response vectors more heavily than individuals with missing data.

The second approach would recognize that propensities have not been perfectly measured and therefore we do not know with certainty which of the three food security classes the sampled individuals fall.

The approach calculates the posterior probabilities of class membership for each of the sampled respondents. Let  $Pr\{\theta_i \leq \tau_1 | \mathbf{x}_i\}$  denote the posterior probability that the  $i$ th individual is food secure (these probabilities would need to be approximated using numerical integration). The weighted average of these posterior probabilities estimates the proportion of the entire population that falls into each of the three classes. The estimate for the proportion of the population that is food secure is:

$$\square Pr\{\text{food secure}\} = \frac{\sum_{i=1}^n w_i Pr\{\theta_i \leq \tau_1 | \mathbf{x}_i\}}{\sum_{i=1}^n w_i}$$

The third approach is the posterior predictive approach. The posterior predictive approach utilizes information from the sampled respondents to predict information about non-sampled units. In the food security survey, we would like to *predict* the food-security index for all individuals in the population. This approach requires one of the two marginal estimation procedures (i.e. the MML or Bayesian approach).

If the population distribution  $F$  were known exactly, then we could calculate exactly the proportion of individuals in each of the three food security classes. The proportions for the three classes are simply  $F(\tau_1)$ ,  $F(\tau_2) - F(\tau_1)$  and  $1 - F(\tau_2)$ . However, the population distribution is not necessarily known exactly.

Suppose that the distribution  $F$  depends on the unknown parameter vector  $\eta$ . The notation  $F_\eta$  is used to remind the reader that the distribution function  $F$  depends on the parameter vector  $\eta$ . There are at least two approaches to estimate the population proportions when the parameter  $\eta$  is unknown.

- The empirical Bayes approach fixes the estimates of  $\eta$  at their maximum likelihood estimates derived from the MML estimation procedure. Let  $F_{\hat{\eta}}$  denote the distribution function calculated using the estimated parameter vector  $\hat{\eta}$ , then  $F_{\hat{\eta}}(\tau_1)$ ,  $F_{\hat{\eta}}(\tau_2) - F_{\hat{\eta}}(\tau_1)$  and  $1 - F_{\hat{\eta}}(\tau_2)$  are the empirical Bayes estimates for the proportion of the population in the three food security classes.
- The fully Bayesian approach requires the posterior distribution of the parameter vector  $\eta$ . Let  $\pi(\eta | \mathbf{X})$  denote this posterior distribution. Then the proportion of the population that is food-insecure without hunger is estimated

$$\square Pr(\text{food insecure}) = \int_{\eta} (F_{\eta}(\tau_2) - F_{\eta}(\tau_1)) d\pi(\eta | \mathbf{X})$$

## SCALE MAINTENANCE

Another decision that must be made when creating a measurement scale is to decide whether the scale will be scaled once using pre-test data, or if it will be rescaled each time the measurement scale is administered (e.g., annually). In either case it is extremely important to determine whether the response behaviors in one administration are similar to those in previous administrations of the survey. The scale items may drift over time. For example, the definition of a “balanced meal” could change over time, and so the behavior of items that target balanced meals may also change over time. If respondents are responding differently to the items, then it is impossible to compare results across administrations.

Probably the simplest way to detect problematic “drifting” items is to perform a DIF analysis. Items that are found to function differently across the two adjacent administrations need to be handled in some way. One way to handle problematic items is to allow them to vary across the two administrations; all other items would be forced to have the same item parameters across the two years, thus allowing us a way to link the two administrations of the survey. The more items in common across the two administrations, the stronger the link will be.

Occasionally it may be determined that some items are no longer useful for detecting food security and the program decides to replace them with new items. As long as some subset of the previous scale’s items are still contained in the measurement scale the two years of data can be linked. The items that are common across the two administrations (that have not drifted) are forced to have the same item parameters, and the parameters for the new items are estimated under that constraint.

## CONCLUDING REMARKS

Item response models are probably the best tool for producing a measurement scale with discretely measured items. However, as with any model, the assumptions must be examined closely. In my own opinion, it is especially important to think about the way in which the cutpoints are defined, and the implications those definitions have on the final results.

While it attempts to cover the most important issues relevant to the development of a measurement scale, this paper does not cover all topics. One important piece that has been left out of the paper is a discussion of how to deal with the complex sampling design. It is not a straightforward problem when inferences are being made about a latent construct like food insecurity. The National Assessment of Educational Progress tackles the problem using plausible values drawn from each individual’s posterior distribution, and a jackknife estimator of the sampling variation. Johnson and Jenkins (2005) take a super-population approach to the problem.

The Food Security program has a lot in common with the National Assessment of Educational Progress. The goals of the two programs may be different, but the methods the two use to reach those goals are quite similar: both use discretely scored items to develop a scale (or scales); both estimate each respondent’s location on the scale and use those estimates to approximate the proportion of the population within a number of classes defined on the propensity scale; and both use complex sampling methods to select their respondents. The Summer 1992 issue of the *Journal of Educational Statistics* is a special issue dedicated to the procedures behind NAEP, and provides more useful information for developing a measurement scale to be used in large-scale

surveys. Technical reports are also produced for each administration of NAEP (e.g., U.S. Department of Educational, Office of Educational Research and Improvement, 2001).

## REFERENCES

- Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B*, 32:283–301.
- Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42:69–81.
- Anderson, S. A. (1990). Core indicators of nutritional state for difficult-to-sample populations. *Journal of Nutrition*, 120:1557–1600.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397–479). Reading, MA: Addison-Wesley.
- Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64:153–168.
- DeVellis, R. F. (1991). *Scale Development: Theory and Applications*, volume 26 of *Applied Social Research Methods Series*. SAGE Publications, Newbury Park, CA.
- Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, 47:7–28.
- Fischer, G. (1987). Applying the principals of specific objectivity and generalizability to the measurement of change. *Psychometrika*, 52:565–587.
- Froelich (2002). Dimensionality of the USDA food security index. Technical report, Iowa State University, Department of Statistics, Ames, IA. Prepared under USDA Economic Research Service cooperative research agreement 43-4AEM-8-80079.
- Ghosh, M. (1995). Inconsistent maximum likelihood for the Rasch model. *Statistics & Probability Letters*, 23:165–170.
- Guttman, L. (1950). *Measurement and Prediction, Studies in Social Psychology in World War II*, volume IV, chapter The Basis for Scalogram Analysis, pages 60–90. University Press, Princeton, NJ.
- Holland, P. W. and Thayer, D. T. (1986). Differential item performance and the Mantel-Haenszel procedure. Research Report 86-69, Educational Testing Service, Princeton, NJ.
- Irtel, H. (1994). The uniqueness structure of simple latent trait models. In Fischer, G. and Lam-ing, D., editors, *Contributions to Mathematical Psychology, Psychometrics, and Methodology*. Springer-Verlag, New York.

- Irtel, H. (1995). An extension of the concept of specific objectivity. *Psychometrika*, 60:115–118.
- Johnson, M. S. (2004). Item response models and their use in measuring food insecurity and hunger. Prepared for the Committee on National Statistics *Workshop of Food Insecurity and Hunger*.
- Johnson, M. S. and Jenkins, F. (2005). A Bayesian hierarchical model for large-scale educational surveys: An application to the National Assessment of Educational Progress. Technical Report RR-04-38, Educational Testing Service, Princeton, NJ.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47:149–174.
- Mislevy, R. and Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models*. Scientific Software, Mooresville, Indiana. [Computer Program].
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. De Gruyter, Berlin.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16:159–176.
- Muraki, E. and Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating scale data*. Scientific Software International, Chicago, Illinois. [Computer Program].
- National Research Council (2005). *Measuring Food Insecurity and Hunger: Phase 1 Report*. Panel to Review U.S. Department of Agriculture's Measurement of Food Insecurity and Hunger, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. Prepublication Copy.
- Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32.
- Nord, M. (1999). Upward bias on food insecurity and hunger prevalence estimates due to measurement error. Working Paper FS-8, U.S. Department of Agriculture, Economic Research Service, Washington, DC.
- Opsomer, J., Jensen, H., Nusser, S., Dringei, D., and Amemiya, Y. (2002a). Statistical considerations for the USDA food insecurity index. Working Paper 02-WP 307, Center for Agricultural and Rural Development, Iowa State University, Ames, IA.
- Opsomer, J., Jensen, H., and Pan, S. (2002b). An evaluation of the USDA food security measure with generalized linear mixed models. Working Paper 02-WP 310, Center for Agricultural and Rural Development, Iowa State University, Ames, IA.
- Patz, R. and Junker, B. W. (1999). Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses. *Journal of Educational and Behavioral Statistics*, 24:342–366.
- Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56:611–630.
- Ramsay, J. O. and Abrahamowicz, M. (1989). Binomial regression with monotone splines: A psychometric application. *Journal of the American Statistical Association*, 84:906–915.

- Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Nielsen & Lydiche, Copenhagen.
- Salzberger, T. (2002). The illusion of measurement: Rasch versus 2-PL. *Rasch Measurement Transactions*, page 882.
- Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.
- Sijtsma, K. and Junker, B. W. (1996). A survey of the methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, 49:79–105.
- Sinharay, S. and Johnson, M. S. (2003). Simulation studies applying posterior predictive checking for assessing fit of the common item response theory models. Technical Report RR-03-28, Educational Testing Service, Princeton, NJ.
- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52:589–617.
- U.S. Department of Education. Office of Educational Research and Improvement (2001). *The NAEP 1998 Technical Report*. NCEES 2001-509, National Center for Educational Statistics, Washington, DC. By N.L. Allen, J. Mazzeo, and others.
- Winsberg, S., Thissen, D., and Wainer, H. (1984). Fitting item characteristic curves with spline functions. Technical Report 84-52, Educational Testing Service, Princeton, NJ.
- h Zhang, J. and Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, 64:213–249.