Item response models and their use in measuring food insecurity and hunger

Matthew S. Johnson

Department of Statistics and Computer Information Systems

Baruch College, City University of New York

Matthew_Johnson@baruch.cuny.edu

July 3, 2004

Abstract

This paper aims to give a general discussion of parametric item response theory models, paying close attention to the Rasch model and its extensions for the analysis of multiple dichotomous and polytomous items. As part of this discussion the paper reviews both the models commonly used in IRT and the procedures utilized to estimate the parameters of these models, and their implications.

After giving a general introduction to IRT models the paper examines the appropriateness of these models for the measurement of food security and hunger. Specifically, the paper examines how appropriate IRT is for the analysis of the food security items by examining a subset of data from the 2002 CPS, and then asks the question of whether or not the propensity measured by the food security items is in fact related to true food insecurity.

Finally, the paper examines how one might classify survey respondents into one of the three food security classes and/or estimate the proportions of individuals in the population that fall into each of these classes.

1 Introduction to item response theory 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 $i \in \{1, \ldots, N\}$ to item $j \in \{1, \ldots, J\}$. Further let $P_{jm}(\theta_i) \equiv Pr\{X_{ij} = m \mid \theta_i\}$, denote the mth *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 θ_i denote the propensity of individual i.

- *Conditional Independence (CI):* Given an individual's propensity θ, the elements of the item response vector for respondent i, $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iJ})^t$, are independent.

- *Monotonicity (M):* $Pr\{X_{ij} > t \mid \theta_i\}$ is a non-decreasing function of an individual's propensity θ_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 θ_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).

2 Examples of item response models

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:

$$\psi_j(\theta) = \log\left\{\frac{P_j(\theta)}{1 - P_j(\theta)}\right\} \tag{1}$$

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

2.1 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)* $P_j(\theta) \equiv Pr\{X_{ij} = 1 \mid \theta\}$ is a non-decreasing function of θ for all items j.

The monotonicity assumption (M) allows us to use the observed item response vector for individual i ($\boldsymbol{x}_i$) as repeated measures of the latent variable θ. In fact, in the dichotomous case, under the conditions U, CI and M the *total score* for individual i, defined as $x_{i+} = \sum_{j=1}^{J} x_{ij}$ has a monotone likelihood ratio in θ (Grayson, 1988; Huynh, 1994). That is

$$\frac{Pr\{x_{i+} > s \mid \theta\}}{Pr\{x_{i+} > r \mid \theta\}} \text{ is increasing in } \theta \text{ for } s > r,$$

and the score x_{i+} consistently orders individuals by their latent variable θ

2.1.1 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 θ and that the slopes of these linear functions are equal across all items.

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

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

The *discrimination* parameter α 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 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. 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 θ. 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. If the probabilities $P_j(\theta_i)$ and $P_j(\theta_{i'})$ are known for two individuals i and i', then the difference between their propensities is:

$$\theta_i - \theta_{i'} = \psi_j(\theta_i) - \psi_j(\theta_i'),$$

which is independent of the item j chosen for comparison. Similarly, if the probabilities $P_j(\theta_i)$ and $P_{j'}(\theta_i)$ are known for two items j and j', then the difference between the item difficulties is

$$\beta_j - \beta_{j'} = \psi_{j'}(\theta_i) - \psi_j(\theta_i)$$

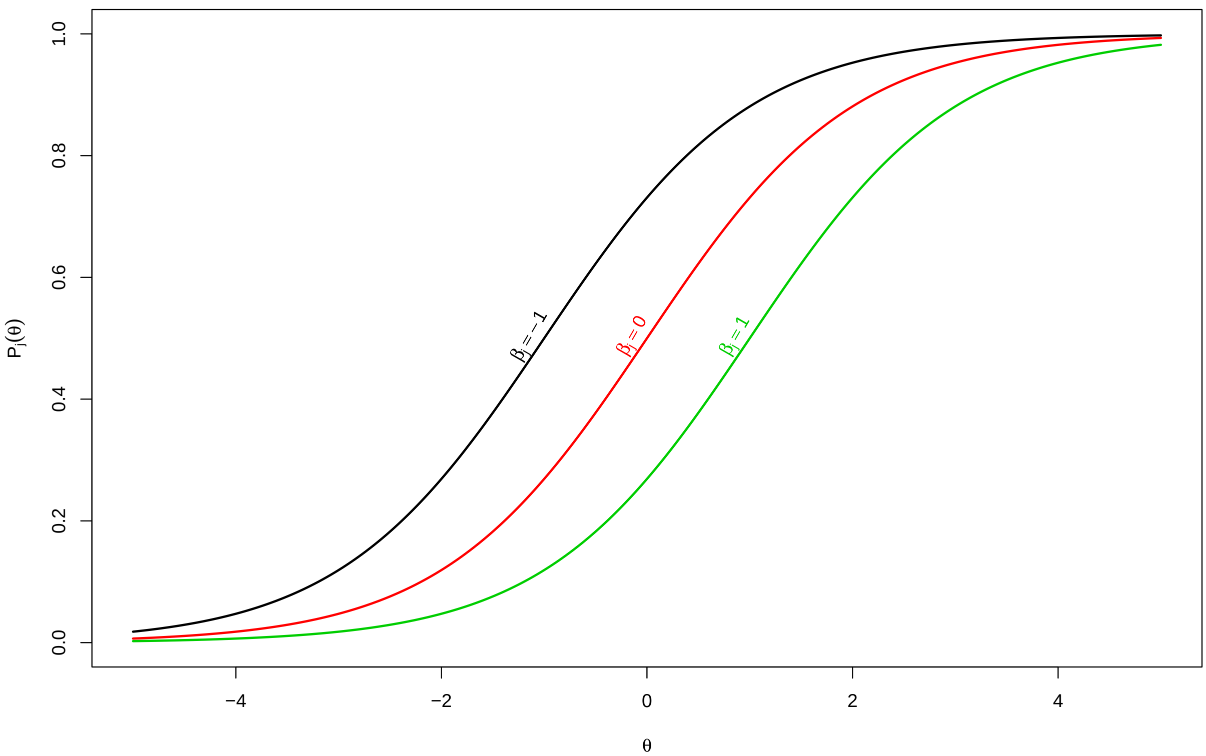

Figure 1: Three examples of the Rasch item response function with slope $\alpha = 1$ and difficulties $\beta = -1,\ 0,\ 1$.

which is independent of the individual i chosen for comparison. Proponents of the Rasch model claim that any method of measurement should be specifically objective, and 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 θ_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. Section 4.1 examines whether or no the Rasch model, which is currently used in the analysis of the food security data, adequately explains the response behaviors of the sampled respondents.

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.

2.1.2 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 which generalizes the Rasch model by allowing the slopes to vary. Specifically the 2PL assumes the following

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

The slope parameter, sometimes called the discrimination of the item, is a measure of how much information an item provides about the latent variable θ. As $\alpha \to \infty$ the item response function approaches a step function with a jump at β_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.

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. The comparison between two individuals i and i'

$$\psi_j(\theta_i) - \psi_j(\theta_{i'}) = \alpha_j(\theta_i - \theta_{i'})$$

depends on the item used for comparison through the discrimination parameter α_j. However, if the discriminations are known then the comparison

$$\frac{1}{\alpha_j}(\psi_j(\theta_i) - \psi_j(\theta_{i'})) = \theta_i - \theta_{i'},$$

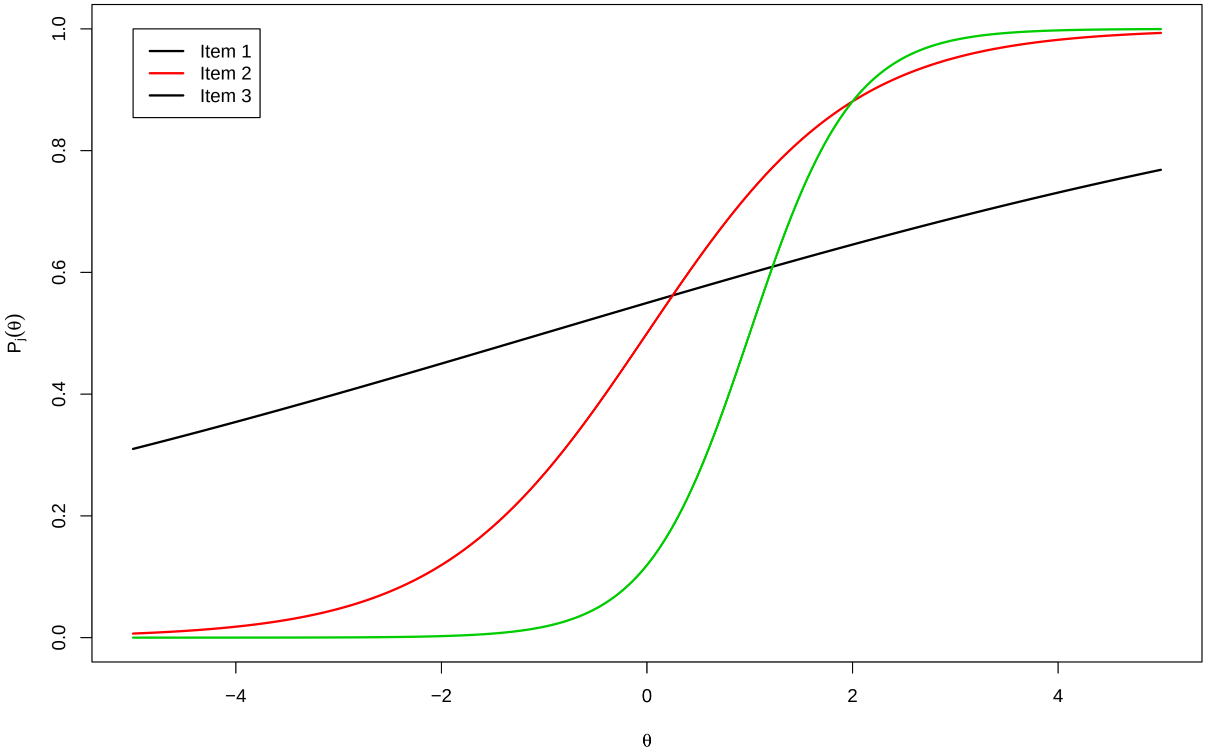

Figure 2: Three examples of the 2PL item response function with slope $\alpha = 0.2,\ 1,\ 2$ and difficulties $\beta = -1,\ 0,\ 1$.

is independent of the item used to compare the two individuals i and i'.

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. Let i, i' and $i^\star$ denote three respondents. Then the function

$$\frac{\psi_j(\theta_i) - \psi_j(\theta_{i^\star})}{\psi_j(\theta_{i'}) - \psi_j(\theta_{i^\star})} = \frac{\theta_i - \theta_{i^\star}}{\theta_{i'} - \theta_{i^\star}}$$

is independent of the item j chosen to make the comparisons.

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 α_j are known the weighted raw score $s_i^\star = \sum_j \alpha_j X_{ij}$ is a minimal sufficient statistic for the propensity θ_i. However, the discrimination parameters are rarely known in advance and must be estimated.

2.1.3 Three parameter logistic model

The response functions $P_j(\theta) \to 1$ as $\theta \to \infty$ and $P_j(\theta) \to 0$ as $\theta \to -\infty$ 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 $\lim_{\theta \to -\infty} P_j(\theta) = 0$ 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

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

The 3PL assumes that the examine knows the correct answer of the item with probability equal to (3) or guesses the item correctly with probability γ_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 θ 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 γ, regardless of the individual's propensity.

2.1.4 Non-parametric IRT models

Many researchers have suggested using the total score x_{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 et al. (1984) on the other hand suggest methods for the estimation of the non-parametric response function $P_j(\theta)$, which utilizes B-splines.

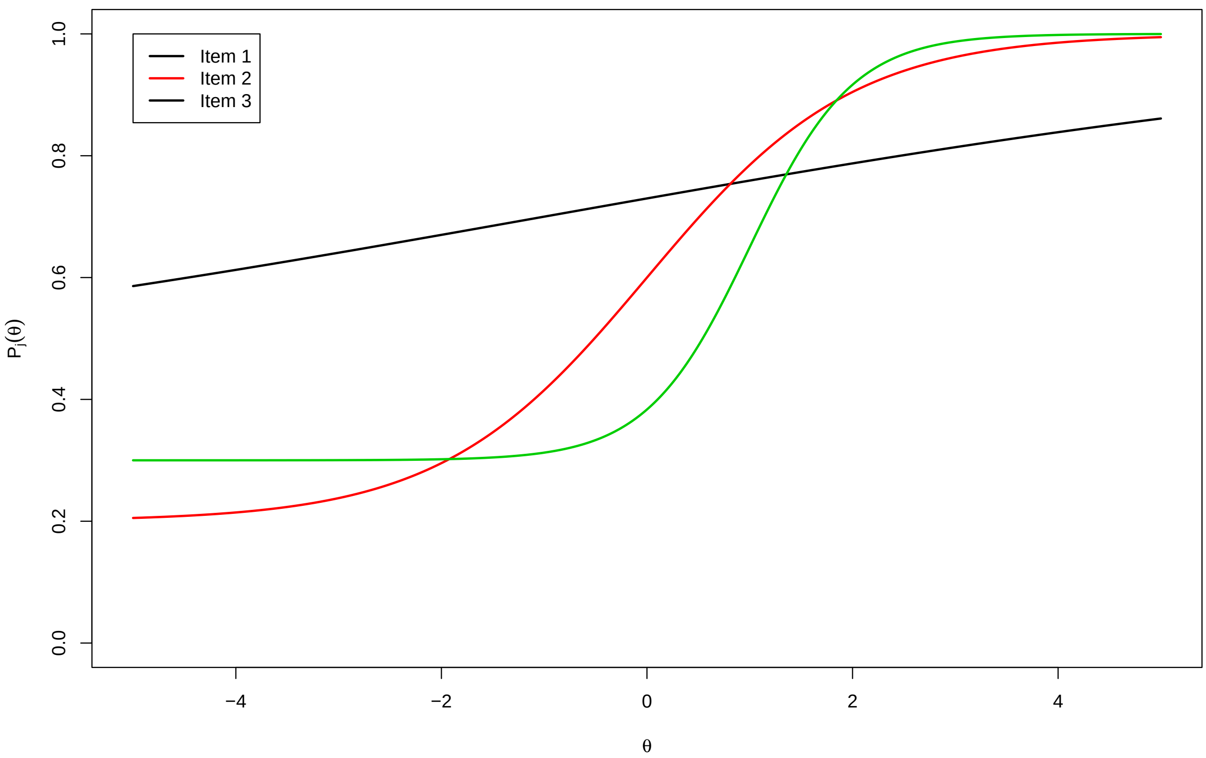

Figure 3: Three examples of the 3PL item response function with slopes $\alpha = 0.2,\ 1,\ 2$, difficulties $\beta = -1,\ 0,\ 1$ and asymptotes $\gamma = 0.4,\ 0.2,\ 0.3$.

Although the B-spline item response model is likely too complicated to use operationally, it can be utilized to examine the appropriateness of the simpler 1-, 2-,and 3-parameter item response models.

2.2 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 4 displays the information curves for a trichotomous item both before (black curve) and after collapsing the upper two categories. The parameters used to create these information

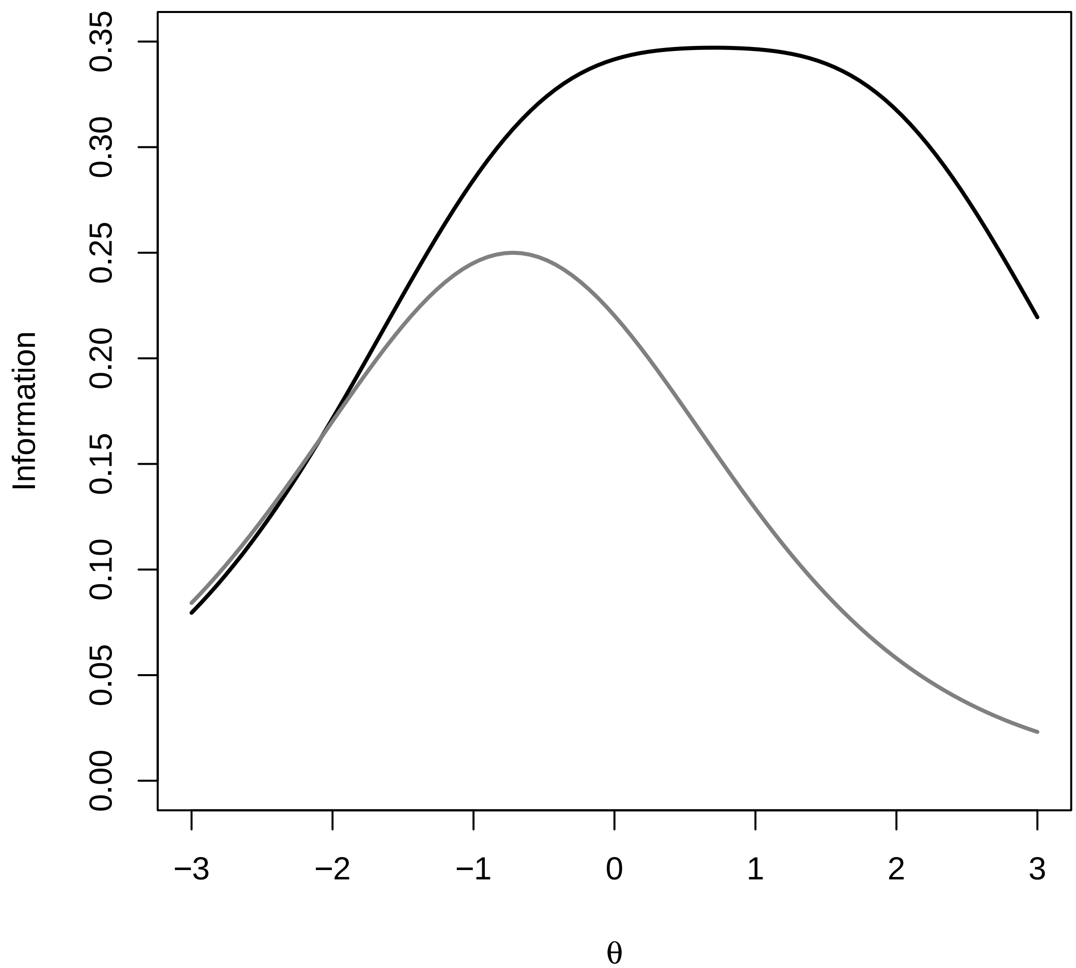

Figure 4: The information about the latent proficiency from a trichotomous partial credit item response before (black curve) and after (gray curve) collapsing the top two categories.

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 θ.

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 $P_{jm}(\theta) = Pr[X_{ij} = m \mid \theta_i]$ is linear in the propensity θ. The descriptions assume that item j is scored with $k_j + 1$ categories, scored 0 to k_j.

2.2.1 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 θ

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

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 β_{jm} are ordered by the category index m, $\beta_{j1} < \beta_{j2} < \cdots < \beta_{jk_j}$.

2.2.2 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 θ:

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

which leads to the following item-category response functions

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

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 et al. (2002a) fit the partial credit model (with constant discrimination parameters $\alpha_{jm} = \alpha_j$) 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.

2.2.3 Rating scale model

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

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

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-category item, $k_j - 1$ dichotomous pseudo-items can be created that are conditionally independent under the model. The mth 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.

3 Estimation

Estimating the model parameters for any item response model requires additional thought about the items and the respondents participating in the survey. Basic estimation techniques for item response models assume that the individuals participating in the survey are independent of one another and that items behave in the same way for all individuals (i.e. there is no differential item functioning present).

There are 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 basic estimation 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 θ_i. Under the assumption of conditional independence the joint probability of the item response vector $\boldsymbol{x}_i$ conditional on θ_i is

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

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

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

The following sections describe the four basic methods for the estimation of item response models.

3.1 Joint maximum likelihood

The joint maximum likelihood (JML) estimation procedure treats both item parameters (e.g. β_j) and propensities θ_i as unknown, but fixed model parameters. Under the JML procedure the $N \times J$ item responses are essentially treated as the observational units in the analysis. The JML procedure estimates the item parameters ($\boldsymbol{\psi}$) and examinee abilities by maximizing $L(\boldsymbol{\psi}, \boldsymbol{\theta}; \boldsymbol{X}) = \prod_i L_i(\theta \mid \boldsymbol{x}_i, \boldsymbol{\psi})$ with respect to $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ simultaneously.

The model is not identified, which means there is no unique solution to the maximization. A unique solution does exist if further constraints are placed on the parameters of the model. For two parameter models like the 2PL, two constraints are necessary: a location constraint, and a scale constraint. The location constraint can be made by constraining either a single propensity or difficulty to some fixed number, or by constraining the average propensity or difficulty to some fixed number (typically zero). The scale constraint can be made by forcing the product of the discrimination parameters to one (i.e. $\prod_j \alpha_j \equiv 1$).

Even after constraints have been placed on the item parameters, the maximization can not be solved analytically and some numerical optimization method must be utilized. The BIGSTEPS program utilized for the analysis of the food security survey performs a joint maximum likelihood procedure that maximizes the joint likelihood function $L(\boldsymbol{\psi}, \theta; \boldsymbol{X})$ with a numerical method called proportional curve fitting.

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.

3.2 Conditional maximum likelihood

Andersen (1970) suggests an alternative method for maximum likelihood estimation of the Rasch model. His method conditions on the vector of raw scores $s_i = \sum_{ij} X_{ij}$, which is a sufficient statistic for the propensities of individuals in the sample:

$$Pr\{\boldsymbol{x}_i \mid \theta_i, \boldsymbol{\psi}, s_i\} = \frac{\exp\{-\sum_j x_{ij}\beta_j\}}{\sum_{\{\boldsymbol{y}:\sum_j y_j = s_i\}} \exp\{-\sum_j y_j \beta_j\}},$$

which does not depend on the value of the individual's propensity θ. The conditional maximum likelihood estimates the item parameters by maximizing the *conditional* likelihood $L(\boldsymbol{\psi} \mid \boldsymbol{X}, \boldsymbol{s}) = \prod_i Pr\{\boldsymbol{x}_i \mid \boldsymbol{\psi}, s_i\}$.

Although Andersen (1970) shows that conditional maximum likelihood estimates for the item difficulties are consistent, an *ad hoc* procedure must be implemented to estimate the propensities of individuals. In addition, the conditional maximum likelihood method only works when there is a simple sufficient statistic like the raw score for the Rasch model. However, as noted earlier more complex IRT models, including the 2PL, do not have simple sufficient statistics.

3.3 Marginal maximum likelihood

Marginal maximum likelihood (MML) takes a different approach to removing the propensities from the likelihood. Unlike joint maximum likelihood estimation techniques, which treat each of the $N \times J$ item responses as separate observational units, the marginal technique treats only the N individuals as the observational units. To accomplish this the MML technique assumes that the propensities are random effects sampled from some larger distribution, denoted $F(\theta)$. 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 a *ordered latent class model*. *Latent variable models* usually refer to models where $F(\cdot)$ is continuous.

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 $\boldsymbol{x}_i$,

$$Pr\{\boldsymbol{x}_i \mid \boldsymbol{\psi}\} = \int_{\Theta} L_i(\theta \mid \boldsymbol{x}_i, \boldsymbol{\psi}) dF(\theta). \tag{6}$$

Taking the product of the probabilities in (6) over individuals i defines the marginal likelihood of the item parameter vector $\boldsymbol{\psi}$

$$L(\boldsymbol{\psi} \mid \boldsymbol{X}) = \prod_i Pr\{\boldsymbol{x}_i \mid \boldsymbol{\psi}\},$$

which is is maximized with respect to the item parameters $\boldsymbol{\psi}$ to derive the MML estimates. Like the JML estimation method, location and scale constraints are required to identify the model. The constraints can either be placed on the mean and standard deviation of the propensity distribution F or on the item parameters.

The propensity distribution F is now a part of the IRT model and care must be taken when choosing the parametric form of F. 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.

In the 2002 Food Security survey approximately 10% of those sampled were administered the entire food security survey. So, if experts agree that the *entire* population of propensities is normally distributed, then the mixing distribution used in MML estimation must account for the non-representative sample. One possible correction, implemented in the analysis section Section 4.1, uses a normal distribution truncated at the 90th percentile, 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 et al., 2002a) to estimate the distribution of food insecurity in all U.S. households.

Another possible way to get around the difficulty of defining a mixing distribution is to assume some non- or semi-parametric form for the mixing distribution. For example, the analysis of the National Assessment of Educational Progress, a large scale educational survey, assumes examinee propensities θ_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 propensity θ is

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

where $\sum_t p_t = 1$. For this distribution of propensities the marginal probability in (6) becomes

$$Pr\{\boldsymbol{x}_i \mid \boldsymbol{\psi}\} = \sum_{t \in \{-4,\ldots,4\}} Pr\{\boldsymbol{x}_i \mid t, \boldsymbol{\psi}\} p_t.$$

The masses p_t are estimated simultaneously with the item parameters $\boldsymbol{\psi}$. Mislevy and Bock (1982) and Muraki and Bock (1997) provide more information on this estimation technique.

In addition to requiring numerical methods to accomplish the maximization of the likelihood, the MML technique also requires numerical integration techniques to approximate the integral in (6).

3.4 Bayesian estimation with Markov chain Monte Carlo

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.

In an IRT analysis of the food security survey items these prior distributions can be utilized to insert expert opinion into the estimation. For example, if experts are 90% sure that the item that asks individuals if they were "ever hungry, but could not afford food" is located above the item that asks if the individual "ever skipped a meal," that information could be incorporated into the model through the prior distribution of the item parameters. The resulting parameter estimates would have the difficulties of these two items switched only if there was overwhelming evidence in the data to support it.

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.

Section 4.1 below implements an MCMC procedure to perform Bayesian estimation of the Rasch and 2PL models for the 2002 Food Security data.

3.5 Estimating the population distribution

The goal of the food security survey is to estimate and report information about the food security of the entire population of individuals. The joint maximum likelihood procedure does estimate the food security propensities for individuals in the sample, but not for the entire population. So, an important question is how to estimate the population distribution from the sampled data.

One suggestion is to use the empirical distribution of the maximum likelihood estimates $\hat{\theta}$ as an estimate of the population distribution. This method suffers in at least two ways. Firstly, with a small number of items, the number of mass points in the empirical distribution is small, even though IRT models assume a continuous distribution for the propensities (Opsomer et al., 2002a). Secondly, the MLEs of the propensity scores are simply *point estimates*. The propensities for the individuals in the survey have not been measured *without error*.

The first problem can be worked around by either assuming some parametric form for the distribution of propensities in the population, and then estimating the parameters of this distribution with the estimates from the joint maximum likelihood procedures, or using kernel or spline smoothing technique to estimate the distribution.

The problem of how to deal with the uncertainty of our imperfect measurements is a little bit trickier under the JML procedure. The marginal estimation techniques however can easily handle estimation of the population distribution. The MML and Bayesian techniques assume that individuals' propensities have been sampled from the population distribution $F(\theta)$. Any parameters associated with this distribution are estimated as part of the MML and/or Bayesian procedure.

In addition to implicitly estimating the population distribution, 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 $\boldsymbol{Y}_i$. If the propensities were known then a logical step in analysis would be to perform a linear regression of θ on $\boldsymbol{Y}_i$.

Although the propensities are latent, this regression can still be performed by assuming the mean of the mixing distribution for individual i is a linear function of the independent variable $\boldsymbol{Y}_i$. In the normal case, this would amount to $\theta_i \sim N(\boldsymbol{Y}_i^t\boldsymbol{\gamma}, 1)$. The estimates of γ from this latent regression can be interpreted directly without having to rely on the estimated propensities of individuals in the sample. Opsomer et al. (2002b) perform such an analysis on food security data.

Clearly the sampling design and/or sampling weights play important roles when estimating the population distribution, and they must be dealt with properly when a marginal estimation procedure is employed.

4 Appropriateness of IRT for the measurement of food security

This section examines the appropriateness of IRT models for the measurement of food security and hunger. I look at the question from two directions. Section 4.1 addresses whether or not a simple IRT model like the Rasch model is appropriate for the analysis of the food security survey questions. Section 4.2 looks at the question of whether or not it is *valid* to treat the propensities derived from an IRT analysis as measures of true food security and hunger.

4.1 Appropriateness of IRT/Rasch for the food security questions

4.1.1 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 θ. For example, under the Rasch model, we implicitly assume that the difficulty parameters are the same for both male and female respondents. If the response behaviors of men and women are not the same, then any estimates of the IRT item parameters and/or population quantities are going to 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 of their propensities θ_i. A Mantel-Haenszel chi-squared test is performed to determine whether or not the score on a given item is independent of the group membership (e.g. gender) of the responding individual within each raw score.

Opsomer et al. (2002b) 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.

4.1.2 Unidimensionality

One of the primary assumptions of the simple IRT models discussed above is that the underlying propensity θ_i is unidimensional. 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.

4.1.3 Missing data

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.

As Opsomer et al. (2002a) 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 propensities 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 et al. (2002a) 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 responses 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.

4.1.4 Adequacy of the Rasch model for the food security data

Whenever any statistical model is utilized the researcher must make sure that assumptions of that model have not been violated; before you make inferences about regression coefficients you need to make sure that the actual relationship between the dependent and independent variables is actually linear. Likewise, when you fit a specific IRT model to multiple discrete responses, you need to make sure the shape of that response model fits the data before making inferences. This section examines the appropriateness of the Rasch model for the food security data.

I extracted food security data from the 2002 CPS. The data contained the responses of 9804 individuals who were administered all of the food security items. These individuals represented 7.89% of the households in the population according to their weights.

The responses of these 9804 individuals were used to approximate the posterior distributions of the model parameters for both the Rasch and the 2PL models. The MCMC procedure assumed flat priors for all item parameters and the mixing distribution assumes that these 9804 respondents were randomly sampled from a normal distribution truncated at the 90th percentile in order to account for individuals who were screened out before receiving the entire food security survey.

The MCMC approximated posterior means for each of the item parameters under the two models appear in Table 1. The common discrimination parameter in the Rasch model is estimated to be 3.50; the geometric mean of the estimated 2PL discrimination parameters, which should be similar to the Rasch discrimination, is 3.20.

Although the rank ordering of the item difficulties is identical under the Rasch and 2PL models, the item discriminations vary quite a bit from item to item under the 2PL model. In particular, the discrimination for the third item, "Couldn't afford a balanced meal," is significantly smaller than eight of the nine other discrimination parameters. Only the difference with the last item is non-

	Rasch	2PL	
Description	Difficulty	Discrimination	Difficulty
Worried food would run out	0.89	3.03	0.79
Food bought didn't last	1.18	3.24	1.16
Couldn't afford balanced meal	1.34	1.94	1.16
Adult cut size or skipped meals	1.37	3.17	1.35
Frequency in the past twelve months	1.51	3.06	1.49
Ate less than felt they should	1.40	3.64	1.41
Hungry but didn't eat	2.02	4.82	1.96
Respondent lost weight	2.28	4.16	2.22
Adult didn't eat for a whole day	2.47	3.56	2.46
Frequency in the past 12 months	1.93	2.36	1.85

Table 1: Item parameter estimates derived from a Bayesian analysis of the ten adult and household food security items.

significant. The discrimination parameter for the item, "Hungry, but didn't eat," is significantly larger than the discrimination parameters of the nine other items.

The Bayesian analysis of the data appears to suggest that there is statistical evidence, at least in the 2002 data, that the Rasch model does not adequately explain the food security survey data. And as noted earlier, items with large discrimination parameters carry the most information about the underlying propensity θ. In the analysis of the 2002 food security data the item that asks individuals if they were ever hungry but didn't eat has the highest estimated discrimination parameter. That item should carry more weight when calculating a food insecurity index.

Figure 5 displays the MCMC approximated posterior densities for two individuals in the food security survey for the 2PL analysis. The first individual, represented by the black curve, responds "Yes" to the first six food security items, and therefore is classified as hungry. The second individual, represented by the gray curve, responds "Yes" to only five of the first seven items; only the balanced meal question and the number of times an adult skipped a meal in the last twelve months were negative responses. The two densities are nearly indistinguishable, yet the individual

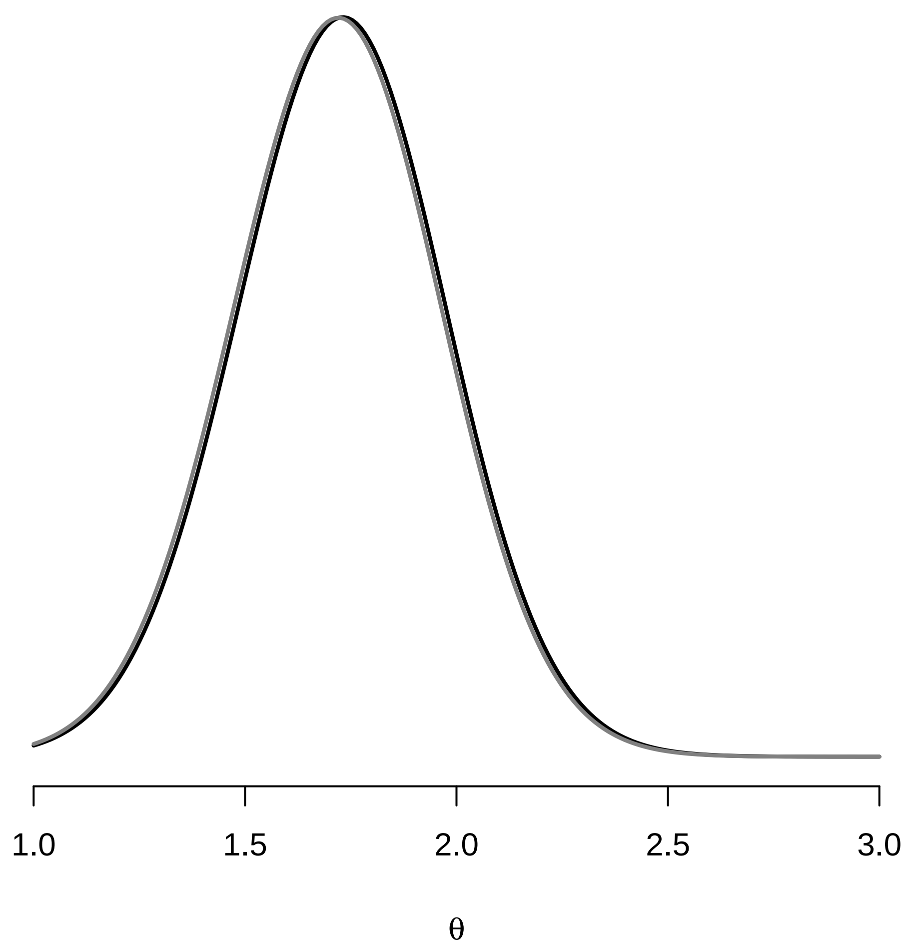

Figure 5: The posterior distributions for two individuals from the 2PL analysis of the 2002 food security data. The solid black curve represents the posterior density of an individuals who answered “Yes” to the first six questions. The gray curve represents an individual who answered “Yes” to only five of the first seven items.

represented by the black curve is classified as food insecure with hunger and the second individual is classified as food insecure without hunger.

4.2 Validity of the IRT propensity as a measure of food security and hunger

The question of validity of the IRT propensity as a measure of food security and/or hunger really comes down to whether or not the latent construct measured by the items is associated with true food-security and/or hunger. Clearly if true food-security is unrelated to the construct measured by these 10 (or 18) food-security items, then it really does not matter how well the item responses and the construct they measure adhere to any specific IRT model.

The only way to be sure that the results from an IRT analysis of the food security items are an appropriate way to measure food-security is to perform some sort of *validity study*. In educational testing validity studies are carried out to determine whether or not college entrance exams actually do help predict future success in college.

A validity study for the food-security program would likely be a difficult task. One such validity study might require monitoring a number individuals over a twelve month period to determine if they were actually food secure, food insecure, or hungry (because they could not afford enough food) at some point during that twelve month period. After that twelve month period the individuals in the validity sample would complete the food-security questionnaire. And finally, an analysis would be performed to determine what, if any, relationship exists between the propensity θ_i measured by the survey items and the true food security status of the individuals in the validity study.

Let $Y_i = 0,\ Y_i = 1,$ and $Y_i = 2$ denote the realizations that individual i was actually food secure, food insecure, or food insecure with hunger respectively. Ideally, there would be a perfect monotone relationship between true food-security/hunger variable Y_i and the propensity θ_i measured by the survey questions. If there was a perfect monotone relationship between the two, then there exists cutpoints τ_1 and τ_2 such that:

$$Y_i = \begin{cases} 0 & \text{if } \theta \leq \tau_1 \\ 1 & \text{if } \tau_1 < \theta \leq \tau_2 \\ 2 & \text{if } \theta > \tau_2 \end{cases} \tag{7}$$

If (7) holds, then Y_i could be treated as a trichotomous (or two dichotomous) Guttman-style item(s) in an IRT analysis with the ten (or 18) food security items to estimate the cutpoints τ_1 and τ_2. These cutpoints would then help to 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.

More than likely the relationship between Y_i and θ_i is a probabilistic relationship, where the distribution of the propensity θ_i depends on the food security and hunger status Y_i of individual i. Let $F(\theta \mid Y = y)$ denote the distribution of the propensity θ given the individuals' food security status is $Y = y$. Hopefully the distributions are *stochastically ordered* so that

$$F(t \mid Y = 0) \geq F(t \mid Y = 1) \geq F(t \mid Y = 2) \tag{8}$$

for all values t. If the propensities are normally distributed within each of the food security groups with means μ_0, μ_1, and μ_2 and standard deviations σ_0, σ_1, and σ_2, where $\mu_y = E[\theta \mid Y = y]$ and $\sigma_y^2 = V(\theta \mid Y = y)$, then the stochastic ordering property in (8) 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 Y_i and the ten food security item responses $X_{i1}, X_{i2}, \ldots, X_{i10}$ have been observed, an analysis would be performed to determine whether there is statistical evidence supporting either of the conditions described above in (7) or (8).

5 Classifying individuals with cutscores

A childless household is classified as food-insecure if the household responds affirmatively to three or more of the ten adult and household referenced items; the household is classified as food-insecure with hunger if the household responds affirmatively to at least six of the ten items (Nord et al., 2002). The raw score is used as the cutoff.

One of the problems I see with this method is that it cannot handle individuals with missing responses. For example, how would you classify an individual who responds to only four of the ten items, but responds positively to those four items? I suppose it might depend on which four items the individual responds to. What about an individual that answers "Yes" to five of eight of the items?

Another problem with the method is that the raw scores are only valid proxies for the estimated propensities if the Rasch model is valid. As the previous section suggests, the Rasch model may not be valid for the food security data.

However, item response models can be utilized to set cutscores on the latent propensity scale, no matter which IRT model is utilized. The following sections discuss some ways in which IRT can be used to estimate the number of individuals in each of the food security classes. Section 5.1 suggests three methods for estimating the population proportions when the cutpoints in (7) are known. Section 5.2 reviews methods for the approximation of cutpoints from expert opinions in the absence of a validity study. And Section 5.3 applies some of these methods to the 2002 data.

5.1 Estimating the proportion of the populations in each hunger class

Suppose that a validity study was performed and the cutoffs τ_1 and τ_2 in (7) were estimated, then 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 here.

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

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

where $I\{\tilde{\theta}_i\}$ is the function indicating whether or not the estimated propensity for individual i is below the first cutpoint.

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 $Pr\{\theta_i \leq \tau_1 \mid \boldsymbol{x}_i\}$ denote the posterior probability that the ith individual is food secure. 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:

$$\widehat{Pr}\{\text{food secure}\} = \frac{\sum_{i=1}^{n} w_i Pr\{\theta_i \leq \tau_1 \mid \boldsymbol{x}_i\}}{\sum_{i=1}^{n} w_i}$$

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

- The empirical Bayes approach fixes the estimates of $\boldsymbol{\eta}$ 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 $\hat{\boldsymbol{\eta}}$, 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 $\boldsymbol{\eta}$. Let $\pi(\boldsymbol{\eta} \mid \boldsymbol{X})$ denote this posterior distribution. Then the proportion of the population that is food-insecure without hunger is estimated

$$\widehat{Pr}(\text{food insecure}) = \int_{\boldsymbol{\eta}} (F_{\boldsymbol{\eta}}(\tau_2) - F_{\boldsymbol{\eta}}(\tau_1)) d\pi(\boldsymbol{\eta} \mid \boldsymbol{X})$$

5.2 Approximating cutpoints with expert opinions

If a validity study for the food security survey is impossible, then some other method must be used to set the cutpoints. The current methodology classifies all individuals who affirm at least six of the items as hungry, but as Bavier (2003) notes a large proportion of the sample who are classified as hungry respond "No" to the direct question about whether or not they were hungry in the last twelve months but did not eat. The operational definition of hunger can be written mathematically as

$$Pr\{\text{hunger} \mid X_{i+} \geq 6\} = 1,$$

and it also implicitly assumes that

$$Pr\{\text{hunger} \mid X_{i+} < 6\} = 0,$$

This classification rule treats the item response data as the fixed quantity (e.g. the independent variable). It seems more reasonable to assume that the item responses vary by food security status. In regression terms you would treat the item responses as the dependent variables and the food security status as the independent variable in analysis.

Suppose that experts agree that individuals who are hungry will respond affirmatively to at least six of the ten items. Mathematically we can write this,

$$Pr\{X_{i+} \geq 6 \mid i \text{ is hungry}\} = 1 \tag{9}$$

We can rewrite the probability on the left hand side of the equation above

$$Pr\{X_{i+} \geq 6 \mid i \text{ is hungry}\} = \int_{\tau_2}^{\infty} Pr\{X_{i+} > 6 \mid \theta_i\} dF(\theta \mid \text{hunger}),$$

where $F(\theta \mid \text{hunger}) = \frac{F(\theta)-F(\tau_2)}{1-F(\tau_2)}$ is the distribution of propensities for individuals who are food-insecure with hunger. The only ways the above equation can equal one are: (a) all items are Guttman-items; or (b) all individuals who are food-insecure with hunger have infinitely large propensities.

One way to overcome this problem is by relaxing the strict requirement in (9) to assume that on average there is a 95% chance (or some other large probability) that households suffering from hunger will respond affirmatively to at least six of the ten items

$$\{X_{i+} \geq 6 \mid i \text{ is hungry}\} = \int_{\tau_2}^{\infty} Pr\{X_{i+} > 6 \mid \theta_i\} dF(\theta \mid \text{hunger}) = 0.95 \tag{10}$$

Inserting the estimated population distribution from a marginal analysis in the above equation for $F(\theta)$ and solving for the cutoff τ_2 would yield an estimate of the cutoff based on the expert opinions.

Other approaches for setting cutpoints using expert opinions include:

- Assuming that food insecure and/or hungry households have a fixed probability of affirming a specific question. For example, experts may want to find the cutpoint τ_2 so that

  $$Pr\{X_{i7} = 1 \mid \text{hunger}\} = 0.95,$$

  where item 7 is the item that asks the individual if they were ever hungry but couldn't eat over the past year.

- Assuming that the minimum probability that a hungry household affirms a specific question is fixed at some large probability. For example, the solution of the equation

  $$P_7(\tau_2) = 0.95$$

  could be used as the cutpoint for classifying food insecure households with hunger.

5.3 Application to the 2002 food security data

Using the Rasch and 2PL estimated population distribution and item parameters I found the cutpoint τ_2 that solves (10). The estimated cutpoint under the Rasch model was 1.91 and the cutpoint estimated with the 2PL was 1.93. That is, 95% of population with propensities greater than 1.91 (1.93) will respond affirmatively to at least six of the food security items under the Rasch (2PL) model.

Using the empirical Bayes estimator for the percentage of individuals whose propensity is greater than this cutoff yields estimates of 2.8% and 2.7% under the Rasch and 2PL models respectively. Relaxing the 95% requirement down to 50% produces estimated cutoffs of 1.31 for the

Rasch model and 1.30 for the 2PL model, which correspond to estimates of 9.5% and 9.7% of the population being food insecure with hunger under the Rasch and 2PL models respectively.

6 Concluding Remarks

Item response theory models are designed for the analysis of multiple discrete item responses, which is exactly what the food security survey contains. However, there are many assumptions that need to be examined before making inferences from these models.

First, one should ask whether or not the Rasch model is adequate for the survey data. I have shown here that the 2002 food security data appears to provide evidence in favor of the 2PL model over the simpler Rasch model, and Froelich (2002) finds that the items appear to be measuring at least a two-dimensional latent trait.

Secondly, even if an item response model is developed that adequately explains the patterns in the food security data, there needs to be a study of how the propensity score measured by these items is related to true food security and/or hunger. If this study is performed, it is relatively simple to develop classification and estimation techniques to estimate the proportion of the population in each of the food security classes.

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B*, 32:283–301.

Anderson, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42:69–81.

Bavier, R. (2003). Critique of hunger prevalence estimates associated with the USDA food security measure.

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.

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 genearlizability 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.

Grayson, D. A. (1988). Two group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, 53:383–392.

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.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent bernoulli random variables. *Psychometrika*, 59:77–79.

Irtel, H. (1994). The uniqueness structure of simple latent trait models. In Fischer, G. and Laming, 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.

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, Berline.

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].

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.

Nord, M., Andrews, M., and Carlson, S. (2002). *Household Food Security in the United States, 2002*. Food Assistance and Nutrition Research Report Number 35, Food and Rural Economics Division, Economic Research Service, U.S. Department of Agriculture.

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 teh USDA food security measure with genearlized 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). *Probablistic models for some intelligence and attainment tests*. Nielsen & Lydiche, Copenhagen.

Salzberger, T. (2002). The illusion of measurement: Rasch vesus 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.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52:589–617.

Winsberg, S., Thissen, D., and Wainer, H. (1984). Fitting item characteristic curves with spline functions. Technical Report 84-52, Educational Testing Service, Princeton, NJ.

Zhang, J. and Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, 64:213–249.