Food and Agriculture Organization
of the United Nations

INTRODUCTION TO ITEM RESPONSE THEORY APPLIED TO FOOD SECURITY MEASUREMENT

Basic Concepts, Parameters and Statistics

VOICES
— of the —
HUNGRY

INTRODUCTION TO ITEM RESPONSE THEORY APPLIED TO FOOD SECURITY MEASUREMENT

Basic Concepts, Parameters and Statistics

By
Mark Nord

Food and Agriculture Organization of the United Nations

Rome, 2014

Recommended Citation:

Nord, M. 2014. Introduction to Item Response Theory applied to Food Security Measurement: Basic Concepts, Parameters, and Statistics. Technical Paper, FAO, Rome.
(available at http://www.fao.org/economic/ess/ess-fs/voices/en)

Information on the author:

Mark Nord consults with FAO on experiential food security measurement methods. He recently retired from the United States Department of Agriculture's Economic Research Service, where he led that agency's work on domestic food security measurement, monitoring, and related research for 15 years.

This paper was commissioned by FAO through the project Voices of the Hungry (VoH). VoH collects information on food insecurity (restricted food access) as experienced by individuals in over 140 countries using the Food Insecurity Experience Scale (FIES). The project also aims to assist countries interested in adopting the FIES as part of national food security monitoring efforts.
For further information please see: http://www.fao.org/economic/ess/ess-fs/voices/en/

Contents

Abstract

The single-parameter logistic item response theory (IRT) measurement model (commonly known as the Rasch model) provides a theoretical base and a set of statistical tools to assess the suitability of a set of survey items for scale construction, create a scale from the items, and compare performance of a scale in various populations and survey contexts. It has been used widely as the statistical basis for survey-based experiential food security measurement.

An important step in the validation of food security data is to assess the extent to which the data are consistent with assumptions of the measurement model. In data that meet those assumptions, household raw score (the number of items affirmed by the household) is an ordinal measure of the severity of food insecurity in the household, and the household severity parameter is an interval-level measure of severity. Neither of these important measurement traits is certain if model assumptions are not met. The psychometric validation process is important in initial surveys of language groups and culturally distinct subpopulations and in previously untested modes of survey administration (such as self-administration or on-line administration). Once a set of questions has been assessed in a large sample of a population or subpopulation and found to adequately meet assumptions of the measurement model, psychometric assessment in subsequent surveys may not be necessary. However, psychometric assessment can be valuable in surveys with important policy implications to increase confidence in the findings.

This paper presents basic concepts and mathematics underlying the Rasch model and describes the model parameters and statistics commonly used to assess food security survey data. More detailed information on the Rasch model is available in Wright (1977; 1983), Fischer and Molenar (1995); Bond and Fox (2001); Baker (1992) and Hambleton *et al*. (1991), and from the website of the MESA psychometric laboratory at the University of Chicago at www.rasch.org. Information about applications of Rasch methods to the development and assessment of food security scales is available in Hamilton *et al.* (1997a; 1997b), Bickel *et al.* (2000); Ohls *et al.*

(2001), Nord and Bickel (2002), Nord (2002); Nord (2003), and Nord (2012), and on the US Department of Agriculture Economic Research Service website at www.ers.usda.gov/topics/food-nutrition-assistance/food-security-in-the-us.aspx.

Basic Concepts: Item Severity and Household (or Respondent) Severity[1]

An essential characteristic of experiential food security measures is that the items comprising them vary across a wide range of severity of food insecurity. The precise severity level of each item (the "item calibration," or "item parameter" discussed below) is estimated empirically from the overall pattern of response to the scale items by the interviewed households. However, the range of severity of the conditions identified by the items is also intuitively evident from the cognitive content of the items. For example, the item, *Adult did not eat for a whole day,* is a more severe manifestation of food insecurity than is the item, *Adult cut the size of meals*, and the latter indicates a more severe level of food insecurity than does the item, *Worried whether food would run out before we got money to buy more*. These differences in severity are observed in the response patterns of surveyed households. More severe items are less frequently affirmed than less severe items. Moreover, a household that affirms an item of mid-range severity is likely to also affirm all items that are less severe. Similarly, a household that denies an item at mid-range is likely to deny all items that are more severe. These typical response patterns are probabilistic, not universal, but they are generally predominant in good quality data.

The Rasch model (named for the Danish Mathematician, Georg Rasch) formalizes this concept of the severity-ordering of items and provides standard statistical methods to estimate the severity of each item and each household and to assess the extent to which the response patterns observed in a data set are consistent with the severity-order concept. The Rasch model was developed primarily in the educational testing field, where multiple

[1] Food security survey questions and derived measures can be referenced either to the individual respondent or to the household as a whole. In this paper, "household" is used throughout.

correct/incorrect items, varying in difficulty, are used to measure an individual's level of knowledge or skill.[2] More generally, the model can be used to assess the location of an individual or household along a continuum—in the present case, a continuum of the severity of deprivation in the basic need for food—by combining information from multiple dichotomous (yes/no) items that vary as to the point on the continuum that each item uniquely reflects. This corresponds exactly to the character of the food insecurity/hunger measurement construct. There is no commonly used language that describes the entire continuum of food insecurity and hunger. It is a latent trait, i.e., not directly observable. People do not say, "On a scale of 1 to 10, my food insecurity is at level 3." But people do speak readily about specific experiences, such as running out of money for food, and the specific behaviors and conditions that result, such as being forced to cut back on quality or quantity of food. Information about these experiences, behaviors, and conditions can be elicited by well-designed survey questions.

Mathematics of the Rasch Model

A Rasch-model based measure of food insecurity is based on the concept that both the indicator items making up the scale and the households responding to the items can be located on the same underlying continuum of severity of food insecurity. The mathematics of the model posit that the probability of a specific household affirming a specific item depends on the difference between the severity-level of the household and the severity of the item. The single-parameter logistic (Rasch model), on which most food security scales are based, assumes specifically that the log-odds of a household affirming an item is proportional to the difference between the "true" severity level of the household and the "true" severity level of the item. Thus, the odds that a household at severity-level h will affirm an item at severity-level i is:

[2] As a result of the educational testing derivation, item parameters are often called "difficulty parameters" in the IRT literature, and respondent parameters are often referred to as "ability parameters." In the food security literature, these are generally referred to as "item severity parameters" and "respondent severity parameters."

(1) $P/Q = e^{(h-i)}$

where P is the probability of affirming the item, Q is 1-P, that is, the probability of denying the item (thus, P/Q is the odds of affirming the item), and e is the base of the natural logarithms. Solving equation (1) for P, the probability that the household affirms the item, can be expressed as:

(2) $P = e^{(h-i)}/(1+e^{(h-i)})$

or, to simplify computation, as:

(2a) $P = 1/(1+1/e^{(h-i)})$

The severity of an item, then, is the severity-level of households that are just at the threshold of affirming or denying that item. The odds that a household will affirm an item right at the severity level of the household is 1, corresponding to a probability of 0.5. The odds that a household will affirm an item with a severity parameter one unit lower than that of the household is e^1, or about 2.7, corresponding to a probability of 0.73 [i.e., 1/(1+1/2.7)]. The probability that the household will affirm an item two units lower than its own severity measure is 0.88, and for an item three units lower, it is 0.95.

Scale Metrics and Average Item Discrimination

Since it is the difference between the household and item parameters that determines the probability of affirmation, it is clear that the metric of a scale can be transformed by adding a constant to both household and item parameters without changing the character of the scale. That is, the size of the intervals on the scale conveys meaningful information, but the zero point is arbitrary. (In statistical terms, it is an interval-level measure, but not a ratio-level measure.) The U.S. Household Food Securiy Scale adopted a metric for the 18-item scale based on a mean

item parameter of 7 for the 18 items in order to keep all item and household parameters positive (Bickel *et al.* 2000). This results in household parameters that range from about 1.5 to 13. The 8-item Voices of the Hungry Food Insecurity Experience Scale (FIES) uses a metric based on mean zero for the 8 items, a common practice in IRT measures. This results in FIES respondent parameters ranging from about -1.0 to +2.5.

Although the size of the interval on a Rasch scale is inherently meaningful, it can be affected by factors such as random measurement error (statistical "noise") in the item responses that are not fundamental to the measurement construct. To meaningfully compare the severities of items between two surveys, it is, therefore, often convenient to multiply the item severity parameters of one of the scales by a constant so as to equate the dispersion of item scores in the two scales. (Dispersion is usually measured by the standard deviation of the item parameters.) In this case the comparison of item severity parameters between the scales is referred to as a comparison of relative item severities.[3] Mathematically, this scale adjustment is equivalent to fitting the Rasch model as in (1) above, with the addition of a discrimination parameter, k, as follows:

$$(3) \quad P_{h,i}/Q_{h,i}=e^{(k(h-i))}$$

For a scale based on a given set of data, the discrimination parameter is inversely proportional to the standard deviation of the parameters of the items in the scale. This relationship is used to assess how well the items in a survey under assessment discriminate, compared to a standard. If both scales are estimated with a discrimination parameter of 1 (i.e., on a logit, or pure logistic metric), then the ratio of the standard deviation of the items in the test data to the standard deviation of the same items in the standard scale compares the average discrimination of the items in the test data to their average discrimination in the standard. This is essentially a comparison of overall model fit between the two data sources.

[3] Making this adjustment to the dispersion of items in a scale under assessment in order to minimize the difference (or squared difference) between the adjusted item parameters of the scale under assessment and the parameters of corresponding items in the standard is theoretically justified by the assumption that the characteristic most likely to be invariant across subpopulations is the severity of the items.

It is worthwhile to consider at this point what item discrimination indicates, and what the practical consequences are of higher or lower discrimination. The extent to which responses conform to expectations consistent with their severity order indicates the strength of association between the item responses and the latent trait—that is, it indicates the discriminating power of the items. Item discrimination is lower if respondents do not clearly understand the meaning of items, if the meaning of items is understood differently by different types of respondents, if the items are not consistently associated with the latent trait, if respondents are not attentive or not taking the survey seriously, or if interviewers do not properly record responses. The practical result of low item discrimination is high measurement error and thus low measurement reliability. For analysis at the household level, measurement error weakens estimated associations with potential causes and consequences of food insecurity. At the population level, measurement error may also bias estimated prevalence rates, since misclassifications in the tail of a distribution are not symmetric.

In general, then, high item discrimination is a desirable trait in response data. However, item discrimination that is too high may indicate that interviewers were "coaching" respondents or filling in some responses without actually administering the questions.

Previous simulation research by the author (unpublished) has found that differences in item discrimination of up to 20 percent have relatively little effect on measurement reliability or on accuracy of population-level prevalence estimates.

Assessing item discrimination, or overall model fit, by the dispersion of items is only meaningful if there is a standard of comparison that is based on data thought to be of good quality, and that contain many items equivalent to those in the data under asssessment. Lacking such a standard, model fit can be assessed by the likelihood ratio for the fitted model compared with a more general model estimated from the same data (such as the proportion of each item affirmed overall), or by the Rasch reliability statistic. However, both of these measures also

require some experiential basis of comparison to be interpretable, and are not described further here.

Rasch Model Estimation and Household Severity Measures

Software that implements the Rasch model begins with the household-by-item matrix of responses. Maximum-likelihood methods are then used to estimate the item severity parameters and household severity parameters most consistent with the observed responses under the Rasch assumptions.[4] The resulting household parameters are a continuous interval-level measure of the severity of food insecurity in the household. These parameters are appropriate for associative analyses such as correlation and regression, with the caveat that the parameter for households with extreme raw scores (those that denied all items or affirmed all items) cannot be estimated by the Rasch model and may differ considerably across households and data sets.

Assessing Items: Item-Fit Statistics

The Rasch model also provides the basis for "fit" statistics that assess how well each item, each household, and the overall data conform to the assumptions of the measurement model. Two statistics commonly used to assess how well responses to items correspond to the Rasch-model assumptions (or "fit" the model) are "item infit" and "item outfit." These are chi-square-type statistics that compare the misfit of each item with the extent of misfit expected under model

[4] Three different maximum likelihood approaches are commonly used to estimate Rasch parameters from item response data: joint (or unconditional) maximum likelihood (JML), conditional maximum likelihood (CML), which is conditioned on raw score, and marginal maximum likelihood (MML). A full discussion of these methods is beyond the scope of this paper. CML is somewhat more complex to implement, but is preferred for applications such as food security scales that comprise relatively few items. JML is simpler to implement and handles missing responses more readily than CML but overestimates the dispersion of item scores, especially in small item-sets. The overestimation of dispersion biases item-fit statistics and assessments of conditional independence of items (an assumption of the Rasch model). MML can be used, and can fit more flexible models (such as 2-parameter models, in which the discrimination of items differs), but has been little-used to date in the food security field.

assumptions. After item and household parameters have been estimated, the probability of an affirmative response in each cell of the household-by-item matrix is calculated. The infit and outfit statistics are then calculated by comparing the actual responses to the probabilistically expected responses in each cell of the matrix. Infit is an "information-weighted" fit statistic for each item, so that it is sensitive to responses by households with severity scores in the range near the severity level of the particular item.[5] Outfit is sensitive to unexpected responses from households with severities much higher or lower than that of the item—that is, to highly improbable responses (outliers).[6]

Both statistics compare observed deviations of responses from the deviations expected under Rasch assumptions, so the expected values of the statistics are 1. Values above 1.0 indicate items that are less strongly or consistently related to the underlying condition (food insecurity) measured by the set of items. (The statistics should really be called item misfit statistics, as higher values indicate poorer fit.) Such an item will have a disproportionate share of "out-of-order" responses (i.e., affirmative responses by households with severity scores below that of

[5] Item infit is calculated as follows:

$$INFIT_i = SUM\ [(X_{i,h} - P_{i,h})^2] / SUM[P_{i,h} - P_{i,h}^2]$$

where:

SUMs are taken for the item across all non-extreme cases
$X_{i,h}$ is the observed response of household h to item i (1 if response is yes, 0 if response is no);
$P_{i,h}$ is the probability of an affirmative response by household h to item i under Rasch assumptions, given the item calibration and the estimated level of severity of food insecurity in the household.

The expected value of each item's infit statistic is 1.0 if the data conform to Rasch model assumptions. Values above 1.0 indicate that the item discriminates less sharply than the average of all items in the scale.

[6] Item outfit is calculated as the average across households of the squared error divided by the expected squared error:

$$OUTFIT_i = SUM\ [(X_{i,h} - P_{i,h})^2 / (P_{i,h} - P_{i,h}^2)] / N$$

where:

SUM is taken for the item across all non-extreme cases
$X_{i,h}$ is the observed response of household h to item i (1 if response is yes, 0 if response is no);
$P_{i,h}$ is the probability of an affirmative response by household h to item i under Rasch assumptions, given the item calibration and the estimated level of severity of food insecurity in the household;
N is the number of households.

The expected value of each item's outfit statistic is 1.0 if the data conform to Rasch model assumptions. Values above 1.0 indicate a higher than expected proportion of "erratic" responses—affirmative responses to a severe item by households that affirmed few other items or denials of a low-severity item by households that affirmed many other items.

the item or denials by households with severity scores above that of the item). Values of infit and outfit below 1.0 indicate items that are more strongly and consistently related to food insecurity than the average item.

The Rasch model assumes that all items discriminate equally sharply, so fit-statistic values (especially infit) that are far from unity call into question the suitability of the item for use in the scale. As a general rule, infits in the range of 0.8 to 1.2 are considered to be very good. Infits in the range 0.7 to 1.3 are usable, and do not distort measurement substantially, but should be improved for general use (Linacre and Wright, 1994). Infit below 0.7 indicates an item that is strongly associated with the underlying condition measured by all of the items (food insecurity). Including such an item may be acceptable practice, but the information provided by the item is undervalued in the equal-weighted Rasch measure. Infit statistics calculated from small samples should be interpreted cautiously, as sampling errors for the statistic are substantial in samples smaller than about 1,000 cases.

Similar standards may be applied to item outfit statistics, but, in practice, outfit statistics are very sensitive to a few highly unexpected observations. As few as two or three highly unexpected responses (i.e., denials of the least severe item by households that affirm most other items) among several thousand households can elevate the outfit for that item to 10 or 20. Carefully interpreted, outfit statistics may help identify items that present cognitive problems or have idiosyncratic meanings for small subpopulations, but there are no standard cutoffs for assessment.

Assessing Items: Conditional Independence

Another important assumption of the Rasch model is that items are conditionally independent. That is, that all correlations among items result from their common association with the latent trait. Or, put another way, that item responses by households with the same true level of

severity of food insecurity are uncorrelated. This assumption is assessed in response data by comparing observed correlations among items with the correlations expected under model assumptions. Any residual correlation violates the model assumption of conditional independence. Of course, there is always the likelihood of some residual correlation due to random processes. In practice, the assumption of conditional independence is often assessed using principle components factor analysis of the residual correlations, which allows assessment of whether residual correlations are larger than those expected by chance and also provides factor loadings, which provide information on character of the residual correlations. Details and standards of this assessment are beyond the scope of this introduction. Suffice it to say that if two items have high residual correlation, it may be necessary to omit one of them. Or, it may be possible to model them jointly as a 3-category variable if the combined variable meets assumptions of a somewhat more complicated polytomous IRT model.

References

Baker, Frank B. 1992. *Item Response Theory: Parameter Estimation Techniques.* New York: Marcel Dekker, Inc.

Bond, Trevor G., and Christine M. Fox. 2001. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences.* Mahway, New Jersey: Lawrence Erlbaum Associates, Publishers.

Bickel, Gary, Mark. Nord, Christopher Price, William L. Hamilton, and John T. Cook. 2000. *Guide to Measuring Household Food Security, Revised 2000.* USDA, Food and Nutrition Service. Available: www.fns.usda.gov/fsec/files/fsguide.pdf

Fischer, Gerhard H., and Ivo W. Molenaar, eds. (1995). *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer-Verlag.

Hambleton, Ronald K., H. Swaminathan, and H. Jane Rogers (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications, Inc.

Hamilton, William L., John T. Cook, William W. Thompson, Lawrence F. Buron, Edward A. Frongillo, Jr., Christine M. Olson, and Cheryl A. Wehler. 1997a. *Household Food Security in the United States in 1995: Summary Report of the Food Security Measurement Project.* Washington DC: Office of Analysis, Nutrition, and Evaluation, Food and Nutrition Service, United States Department of Agriculture. Available: www.fns.usda.gov/sites/default/files/SUMRPT.PDF

Hamilton, William L., John T. Cook, William W. Thompson, Lawrence F. Buron, Edward A. Frongillo, Jr., Christine M. Olson, and Cheryl A. Wehler. 1997b. *Household Food Security in the United States in 1995: Technical Report.* Washington DC: Office of Analysis, Nutrition, and Evaluation, Food and Nutrition Service, United States Department of Agriculture. Available: www.fns.usda.gov/sites/default/files/TECH_RPT.PDF

Linacre, J.M., and B.D. Wright. 1994. "Reasonable Mean-Square Fit Values." Rasch Measurement Transactions *8(3):370.* Available: www.rasch.org/rmt/rmt83.htm

Nord, Mark. 2002. *A 30-Day Food Security Scale for Current Population Survey Food Security Supplement Data.* E-FAN No. 02015. USDA, Economic Research Service. Available: http://webarchives.cdlib.org/sw1tx36512/http:/www.ers.usda.gov/Publications/efan02015/

Nord, Mark. 2003. "Measuring the Food Security of Elderly Persons," *Family Economics and Nutrition Review* Vol. 15, No. 1, pp. 33-46. Available: www.cnpp.usda.gov/Publications/FENR/V15n1/fenrv15n1.pdf

Nord, Mark. 2012. *Assessing Potential Technical Enhancements to the U.S. Household Food Security Measures,* Technical Bulletin No. TB-1936, USDA, Economic Research Service. Available: www.ers.usda.gov/publications/tb-technical-bulletin/tb1936.aspx.

Nord, Mark, and Gary Bickel. 2002. *Measuring Children's Food Security in U.S. Households, 1995-99.* FANRR- 25, USDA, Economic Research Service. Available: http://webarchives.cdlib.org/sw1tx36512/http:/www.ers.usda.gov/Publications/fanrr25/

Ohls, James, Larry Radbill, and Allen Schirm. 2001. *Household Food Security in the United States,1995 -1997: Technical Issues and Statistical Report.* Prepared by Mathematic Policy Research, Inc., for USDA, Food and Nutrition Service. Available: www.fns.usda.gov/household-food-security-united-states-1995-1997

Wright, B. D. 1977. *Solving Measurement Problems with the Rasch Model.* Mesa Psychometric Laboratory, the University of Chicago, College of Education, Chicago, IL. Available: www.rasch.org/memos.htm

Wright, B. D. 1983. Fundamental Measurement in Social Science and Education. Mesa Psychometric Laboratory, the University of Chicago, College of Education, Chicago, IL. Available: www.rasch.org/memos.htm

Contact:
Statistics division (ESS)
The Food and Agriculture
Organization of the United Nations
Viale delle Terme di Caracalla
00153 Rome, Italy
http://www.fao.org/economic/ess/ess-fs/voices/en

I3946E/1/07.14