

Advances in Developing Country Food Insecurity Measurement

Comparison of a Qualitative and a Quantitative Approach to Developing a Household Food Insecurity Scale for Bangladesh^{1,2}

Jennifer Coates,³ Parke E. Wilde, Patrick Webb, Beatrice Lorge Rogers,
and Robert F. Houser

Gerald J. and Dorothy R. Friedman School of Nutrition Science and Policy, Tufts University,
Boston, MA 02111

ABSTRACT This paper compares a qualitative and a quantitative (Rasch) method of item assessment for developing the content of a food insecurity scale for Bangladesh. Data are derived from the Bangladesh Food Insecurity Measurement and Validation Study, in which researchers collected 2 rounds of ethnographic information and 3 rounds of conventional household survey data between 2001 and 2003. The qualitative method of scale development relied on content experts and respondents themselves to evaluate household food insecurity items generated through ethnographic research. The quantitative method applied the Rasch model to assess the fit of the same items using representative survey data. The Rasch model was then used to test for differential item functioning (DIF) across diverse demographic and geographic subgroups. The qualitative assessment flagged and discarded 10 items, leaving 13. The Rasch assessment of infit and outfit flagged 3 items, and the Rasch DIF test discarded another 10 items, leaving a total of 10 items in the Rasch-derived scale. The 2 scales contained 8 of the same items. The qualitatively and quantitatively derived scales were highly correlated ($r = 0.96$, $P < 0.01$), and the 2 methods located 90% of households in the same food insecurity tercile. This convergence lends added confidence to the use of either scale for identifying food-insecure households in different regions of Bangladesh. Multiple methods should continue to be applied in a systematic and transparent way to lend additional credence to the results when they converge and to pinpoint directions for further clarification where they do not. *J. Nutr.* 136: 1420S–1430S, 2006.

KEY WORDS: • *food insecurity* • *mixed-methods* • *Bangladesh* • *qualitative methods*
• *quantitative methods*

Mixed-method studies, employing quantitative and qualitative techniques, are increasingly valued in the drive to address pressing nutrition and food security problems. (1,2). The food insecurity measurement literature that has been published over the last several years has illustrated the value of combining

different types and sources of data to triangulate the validity of a scale. And yet, the practical challenges of integrating qualitative and quantitative information have often been overlooked.

Anyone who has attempted to develop and validate a scale will recognize that the process, in its ideal form, is a happy marriage between theory generation and empirical confirmation, one that iterates between qualitative and quantitative techniques to construct and test items and to assess their appropriateness for the measure. Quantitative and qualitative methods have unique, complementary strengths and different weaknesses, which means that, through triangulation, they can increase the level of confidence in the inferences drawn from a measure's results (3,4). For instance, quantitative methods are traditionally seen as providing complementary breadth to the depth of insight generated by a qualitative approach.

However, as with many marriages, the execution of the relationship is not always smooth, particularly when different approaches to the same question yield conflicting information. This issue is particularly salient when the construct of interest, in this case household food insecurity, is considered an unobservable, latent trait for which there is no gold standard against which the "truth" can be objectively verified. If incongruities cannot be reconciled, then, depending on the philo-

¹ Published in a supplement to *The Journal of Nutrition*. This publication was made possible through support provided to the Food and Nutrition Technical Assistance (FANTA) Project from the Office of Health, Infectious Disease and Nutrition of the Bureau for Global Health and the Office of Food for Peace of the Bureau for Democracy, Conflict and Humanitarian Assistance at the U.S. Agency for International Development, under terms of Cooperative Agreement HRN-A-00-98-00046-00 awarded to the Academy for Educational Development (AED). The opinions expressed herein are those of the author(s) and do not necessarily reflect the views of the U.S. Agency for International Development. Guest Editors for this publication were Jennifer Coates, Edward A. Frongillo, Anne Swindale, Beatrice Lorge Rogers, Patrick Webb, and Paula Bilinsky. Guest editor disclosure: Jennifer Coates received compensation from AED for additional noneditorial support for the management of the supplement publication; Edward A. Frongillo has no relationships to disclose; Anne Swindale is Deputy Director of the Food and Nutrition Technical Assistance Project and is employed by the Academy for Educational Development; Beatrice Lorge Rogers has no relationships to disclose; Patrick Webb has no relationships to disclose; and Paula Bilinsky is an employee of AED.

² Author disclosure: J. Coates, P. Webb, B.L. Rogers, see above; P.E. Wilde and R.F. Houser, no relationships to disclose.

³ To whom correspondence should be addressed. E-mail: jennifer.coates@tufts.edu.

sophical leanings of the researchers, the conflicting outcomes are often rationalized away in favor of the qualitative and quantitative model of choice.

The United States household food security measurement effort offers 1 such example of the challenge of reconciling divergent results of mixed methods. The 18-question scale that is currently used to assess U.S. household-level food insecurity was intended to reflect an accumulated body of theory and qualitative evidence that gave definition to the boundaries of the food insecurity construct and suggested which domains (uncertainty and worry, inadequate quality, insufficient quantity, and social unacceptability) should be part of U.S. measures of food access (5,6). Additional ethnographic research also suggested that certain population subgroups, for instance the elderly (7–11) and Hispanics (12,13), express their experience differently from one another and from the nonelderly white women on which the original ethnographic research was conducted.

The statistical model that was used to guide the development of the U.S. Household Food Security Survey Measure (US HFSSM)⁴ is the single-parameter Rasch model. This model has its roots in item response theory, which has been applied to educational testing and is often used to develop measures of “ability” or “knowledge.” The Rasch model describes a set of rules and conditions that must be met to achieve important fundamental measurement properties. Rasch model proponents believe that the data should conform to these assumptions in order for the unobservable construct to take shape statistically and have predictable properties as a scale. The U.S. Department of Agriculture (USDA) subjected potential scale items, suggested by previous qualitative work, to evaluation using the Rasch model. However, based on Rasch results, items representing conceptually important elements of the construct (e.g., socially unacceptable coping strategies) were excluded from the final product (14,15). In this instance of conflicting information, it appears that aspects of content validity were sacrificed to uphold the specifications of a favored model.

Was this the right decision? The purpose of the current paper is not to judge the choices of the U.S. scale developers, per se, but to suggest an alternative approach using the example of household food insecurity measurement in Bangladesh. As Brewer (3) points out, “the multimethod approach to... contradictions is to accept the fact that no method measures perfectly and to exploit the fact that multiple measurement offers the chance to assess each method’s validity in the light of other methods.” This paper compares the results of a qualitative process of item assessment to a quantitative (Rasch-based) approach. In doing so, the paper seeks not only to assess the level and type of agreement between the results but, where they diverge, to suggest limitations in the methods themselves that should be considered in using one or a combination of approaches in future.

DATA AND METHODS

This paper is 1 in a series of research papers to emerge from the Bangladesh Food Insecurity Measurement and Validation Study (FIMVS), a 3-year Food and Nutrition Technical Assistance (FANTA)-funded initiative to explore suitable processes for constructing household food insecurity scales for use in a developing country context. A complete account of the data collection methods has been reported

elsewhere (16). In brief, the Bangladesh FIMVS was conducted in 4 phases. The first phase used qualitative methods to explore the relevance of the U.S. experience of food insecurity to Bangladesh and to develop a prototype food insecurity instrument. During the second phase, the food insecurity instrument and a comparator indicator questionnaire (that collected data on demographics, anthropometry, morbidity, income sources, assets, dietary intake, and expenditure) were administered to men and women in 600 households in rural areas and small urban centers across 3 regions of Bangladesh. The third phase consisted of additional qualitative work to further assess the appropriateness of items using food security experts and respondent input, and the fourth phase resurveyed the original 600 households to assess household level food insecurity dynamics. The survey rounds were implemented approximately 1 year apart over the period 2001–2003. The data from this paper are drawn from the first 3 of the 4 phases.

The present analysis proceeded in 4 steps: 1) confirming the relevance of U.S. food insecurity concepts in Bangladesh and generating an initial pool of scale items; 2) assessing the items qualitatively using a content expert group and respondent cognitive debriefing; 3) assessing the same items quantitatively using the single-parameter Rasch model to test for model fit across the population and among demographic and geographic subgroups; and 4) comparing the results of the qualitative and quantitative assessments.

Assessment of relevance of U.S. food insecurity experience to Bangladesh and candidate item generation

The first step in building the case for validity involves clearly defining the concept to be measured and selecting a pool of items believed to be relevant to the construct’s domains (17). Thus, the objective of the first phase of the FIMVS study was to evaluate the extent to which the elements of the food insecurity experience, as described in the U.S. literature, were also evident in Bangladesh. This phase also involved determining whether there were other domains of the experience that were Bangladesh specific and identifying locally relevant behaviors and perceptions, related to these domains, as potentially suitable scale items.

Evidence of the food insecurity experience in Bangladesh was amassed through a comprehensive review of the Bangladesh literature with a bearing on the subjective experience of food deprivation. Following the review, an anthropologist was contracted to conduct ethnographic interactions with villagers in one rural part of the country to explore the social meanings and patterns of meal taking, dietary practices, coping strategies, and words used to express hunger and food insecurity. As shown in **Table 1**, 38 items were identified from these ethnographic data as candidates to be assessed further. All of these items related to 1 of the 4 domains (uncertainty and worry, inadequate quantity, insufficient quality, and social unacceptability) that had been previously described in the U.S. literature.

Qualitative assessment of candidate items

Content experts. The qualitative assessment of candidate items used 2 primary data sources. The first, a “content expert panel” (18, 19), is the most widely advocated approach in psychology and educational testing for determining which of many potential items best represent the intended content of the scale. In the present study, these experts were food security researchers from Tufts University, representatives of the USDA that had been involved in developing the US HFSSM, the directors of a Bangladeshi-owned survey research firm, Data Analysis and Technical Assistance, Ltd. (DATA), with over 30 y of combined experience in food security research, and a select group of enumerators with extensive field experience in addition to higher-level degrees. Experts were selected to represent a diversity of viewpoints and specializations and to contribute their intimate familiarity of local conditions to the process.

The process of item review occurred over several rounds of participatory and consensus-building discussion. Based on Messick’s (17,19) definition of the content aspect of construct validity, the experts were asked to consider the following criteria in reviewing items: 1) Do these items relate to 1 of the 4 underlying food insecurity domains? (Relevance to construct.) 2) Do items, as a group, represent a range of

⁴ Abbreviations used: DATA, Data Analysis and Technical Assistance, Ltd.; DIF, differential item function; FANTA, Food and Nutrition Technical Assistance; FIMVS, Food Insecurity Measurement and Validation Study; IRT, Item Response Theory; USDA, United States Department of Agriculture; US HFSSM, United States Household Food Security Survey Measure.

TABLE 1

Candidate food insecurity scale items and their abbreviations by hypothesized conceptual domain

| | Item description | Item abbreviation | Hypothesized feature of food insecurity |
|----|--|----------------------|---|
| | Did behavior occur in last 12 months due to lack of resources: | | |
| 1 | Family not eat meat as part of an ordinary meal | No meat | Quality |
| 2 | Not give children money for snacks | No kid snacks | Quality |
| 3 | Not purchase snacks for the family | No hh snacks | Quality |
| 4 | Had to eat wheat (or another grain) | Ate wheat | Quality |
| 5 | Not cook <i>bhalo mondo</i> ("rich food") | No rich food | Quality |
| 6 | Ate <i>Mishti Alu</i> (sweet potato) | Ate sweet potato | Quality |
| 7 | Ate <i>Bhatar Mar</i> (rice starch) | Ate rice starch | Quality |
| 8 | Ate <i>Bon Kochu</i> (wild taro) | Ate wild taro | Quality |
| 9 | Ate <i>Shaluk</i> (water lily) | Ate water lily | Quality |
| 10 | Ate <i>Gom Baja</i> (fried wheat) | Ate fried wheat | Quality |
| 11 | Ate <i>Ata Gola Pani</i> (Flour and water gruel) | Ate wheat gruel | Quality |
| 12 | Ate <i>Khud</i> (broken rice) | Ate broken rice | Quality |
| 13 | Children ate <i>Mishti Alu</i> (sweet potato) | Kids ate sweet pot | Quality |
| 14 | Children ate <i>Bhatar Mar</i> (rice starch) | Kids ate rice starch | Quality |
| 15 | Children ate <i>Bon Kochu</i> (wild taro) | Kids ate wild taro | Quality |
| 16 | Children ate <i>Shaluk</i> (water lily) | Kids ate water lily | Quality |
| 17 | Children ate <i>Gom Baja</i> (fried wheat) | Kids ate fried wheat | Quality |
| 18 | Children ate <i>Ata Gola Pani</i> (Flour and water) | Kids ate wheat gruel | Quality |
| 19 | Children ate <i>Khud</i> (broken rice) | Kids ate broken rice | Quality |
| 20 | Not eat square meals | Few square meals | Quantity |
| 21 | Could not eat big fish (for example, carp, hilsha etc.) | No big fish | Quantity |
| 22 | Respondent ate less food | Less food | Quantity |
| 23 | Respondent skipped entire meals | Skipped meals | Quantity |
| 24 | Respondent not eat for an entire day | Not eat for day | Quantity |
| 25 | Children skipped entire meals | Kids skipped meals | Quantity |
| 26 | Main working adult in family skipped entire meals | Work adult skipped | Quantity |
| 27 | Children not eat for an entire day | Kids not eat for day | Quantity |
| 28 | Food stored in the home ran out | Food ran out | Insecurity |
| 29 | Worried about where food would come from | Worried about food | Insecurity |
| 30 | Family purchased rice frequently | Bought rice often | Insecurity |
| 31 | Borrowed money from local moneylenders | Took money loan | Acceptability |
| 32 | Took food (rice, lentils etc.) on credit | Took shop food loan | Acceptability |
| 33 | Borrowed food from relatives or neighbors | Friends food loan | Acceptability |
| 34 | Borrowed food to serve to <i>Attio Shojan</i> or <i>Kutum</i> (guests/ | Loan for guest | Acceptability |
| 35 | Had to seek <i>Kurbani</i> meat (charity meat during Eid) | Sought Kurbani meat | Acceptability |
| 36 | Received or sought <i>Jakat</i> or <i>Fitra</i> (charitable contributions) | Took charity | Acceptability |
| 37 | Not purchase something else to buy food | Forego nonfood | Acceptability |
| 38 | Sold or mortgaged things for food | Sold items | Acceptability |

Source: Tufts FSNP/FANTA data (2001).

severity of the food insecurity problem, from mild to severe insecurity? (Representativity of construct.) 3) Are items worded clearly and unambiguously in English and Bangla? (Technical quality.) 4) Will items be understood the same way across 3 regions of Bangladesh and across various social and demographic groups? (Generalizability to the population.)

Respondent cognitive debriefing. There has been increased recognition by scale developers of the need to garner input from representative respondents about the cognitive process involved in hearing, thinking about, and replying to survey questions (20). One common method used in test development, called the "think aloud method," asks the test takers to narrate their thought processes while composing their written responses (21). This method was considered inappropriate for Bangladesh because individuals to whom questions were administered orally could not narrate their cognitive processes while composing an oral response. A similar approach, applied previously in developing food insecurity scales, is called "cognitive debriefing" (22). This method requires enumerators to inquire about a respondent's item comprehension after the item has been answered. In this case, enumerators performed cognitive debriefings both 1-on-1 with male and female respondents and also later in a focus group setting. In the focus group discussions, the enumerator read a question aloud and solicited information about specific phrases that had been flagged in earlier interviews. Respondent suggestions were sought regarding

improved phrasing and alternative items. This information was used, along with feedback from the expert group, to determine which items should be modified or dropped from the scale.

Quantitative assessment of items using the Rasch model

Rasch and its assumptions. The statistical model used in this study to assess the appropriateness of items for the Bangladesh scale was a type of nonlinear factor-analytic approach called the single-parameter Rasch model. Though this model was used to develop the U.S. Household Food Security Measure, its roots are in psychometrics and item response theory, where it is commonly employed to construct educational tests intended to gauge "ability" or "knowledge" based on an individual's responses to progressively more difficult questions. In the food insecurity literature, the unobservable construct of interest is "household food insecurity," rather than "ability," and the items representing the underlying phenomenon are arrayed along a continuum of "severity" rather than "difficulty."

The Rasch model requires unidimensionality, which assumes that the items on the questionnaire collectively assess a single latent trait along a continuum, in this case, the severity of food insecurity (23). Mathematically, the Rasch model has 2 basic assumptions. The first key assumption is monotonicity (23). As explained by Wilde (24), this means that the probability (expressed as a log-odds) of affirming any

particular item is a simple linear function of a household-specific *food insecurity score* and an item-specific *severity calibration*. With all else held equal, the greater the household food insecurity score, the more likely the household is to respond affirmatively to an item. The more severe an item, the less likely the household is to affirm it. The probability of affirming any item is a function of the difference between the item severity calibration and the true household food insecurity score (24–27). The second key assumption is that of conditional independence, where the likelihood of a household affirming 1 item is independent of its responses to other items (24), conditional on the household's level of food insecurity. Together, these assumptions mean that the probability of affirming any 1 item is the same for households with the same level of food insecurity (24).

According to Hamilton et al. (14), in the U.S. food insecurity measurement initiative, the Rasch model was selected in preference to linear factor-analytic approaches because the Rasch model “more accurately characterizes the covariation among items in the data set than traditional linear factor analysis models” (14 p. 5). More specifically, the dichotomous nature of the scale items violated certain statistical assumptions of linear factor analysis, including the assumption of normally distributed error variance (14). After presenting evidence that the construct was unidimensional, the scale developers chose the Rasch model over other Item Response Theory (IRT) models.

If the model assumptions are upheld by the data, then the Rasch model produces a scale with several desirable properties. For instance, it has the property of additivity, meaning that the measurement units (logits) remain the same distance apart over the entire span of the scale (28). This implies, for instance, that the distance between scores of 8 and 7 is the same as the distance between scores of 2 and 1. The scale can be used in basic mathematical functions such as addition and subtraction, and the arithmetic mean of the scale can be taken to measure the average. Also, the unique Rasch property of parameter separation implies that the severity of an item does not depend on the specific households used in the calibration (28).

In certain statistical modeling approaches (e.g., regression) “fit statistics are used to discover a model that fits the data well” (28). If the assumption is that some Rasch model is correct, however, fit statistics are used to identify those data that do not meet the model requirements and are therefore not useful according to the model. For the purposes of this paper, population-level fit statistics were used first to identify and exclude items that did not meet the model assumptions. Following this step, a differential item function (DIF) procedure was used to identify and exclude items that did not share the same meaning across different demographic and geographic population subgroups.

Model fit. The 2 most common statistics used to assess the degree to which items conform to Rasch model specifications are *infit* and *outfit* statistics. These statistics, which represent the difference between the item performance as expected by the model and the observed performance as informed by household responses, are commonly reported as mean squares: the mean of the sum of the squared standardized residuals for the item. Rasch practitioners typically rely more on the *infit* statistic to diagnose item misfit because this statistic incorporates additional information by weighting more heavily those households' responses that are closest to the item value (29). *Outfit* measures, which are unweighted, are easily influenced by just a few unexpected response patterns.

The expected value of both statistics for each item is 1. Linacre and Wright (30) suggest that a range of 0.8–1.2 is an acceptable deviation from the expected value for high-stakes tests. Because item *infit* and *outfit* <0.80 actually indicate *overfit*, or the redundancy of an item with other items, the developers of the US HFSSM did not concern themselves with the lower bound and instead excluded any item with *infit*s and *outfit*s both higher than 1.2 (14). Based on the U.S. precedent, the same criterion was applied to candidate items in this study.

First, female item responses from the 600 households participating in the first survey round were recoded into dichotomous variables before being fit to the model. Item responses were recoded as “never,” “rarely,” or “sometimes” versus “often” or “mostly,” a coding that was designed to produce a scale that would be more sensitive in detecting the chronically (rather than the episodically) food insecure. A score of 1 indicated greater food insecurity on a particular item, and 0 indicated

less. Though 38 items were generated by the initial phase of qualitative research and tested in the 600-household survey, items that were not applicable to all households in the sample were left out of this particular statistical analysis (these items were retained as candidates in other analyses not presented in this paper, for instance, where only households with children were of interest). For instance, 2 items, related to the traditional Islamic practices of seeking meat donations during the *Kurban* Eid and accepting other kinds of *jakat* or *fitra* (charitable contributions), were not relevant to the non-Muslim minority and were excluded. All child-referenced items were also excluded from the outset because approximately one-third of the sample did not have young children. Three additional items, related to mortgaging productive assets for food, having to eat *shuluk* (water lily), and forgoing some other basic need to obtain food were left out because there was little response variability (items were affirmed by only 2 and 3 households, respectively). In this analysis these items were excluded a priori because of the concern that their presence in the model would skew the severity calibrations and fit statistics of the remaining items. With the Rasch software package WINSTEPS 3.55[®], the remaining 23 items that had been tested in the representative survey were fit to the model, and their *infit* and *outfit* statistics were assessed. Items with *infit* and *outfit* >1.2 were excluded from the scale.

Differential item function. According to fundamental Rasch assumptions, all households with the same food security score have the same probability of an affirmative response to the items (24). In other words, the model expects that item responses are affected only by the latent variable, the severity of the item, and measurement error. Other household characteristics, such as geographic locale, religion, or gender, are associated with the probability of affirmative response to items only through their influence over the household food insecurity score. When an item's severity is affected by such characteristics, even after controlling for the food insecurity score, the item is said to have differential item function (also called “bias”). A test of DIF can be thought of as assessing, item by item, whether the distance between a particular item severity calibration and that of a reference item (typically set to 0) for 1 subgroup is the same as the distance between the item calibration and the reference item for another subgroup while holding constant the household food insecurity score. The null hypothesis is that the 2 distances do not differ significantly from one another. The alternative hypothesis is that they do. A *t* test is used to assess the significance of these differences. Items that demonstrate statistically significant DIF are said to have a different “meaning” across subgroups. There are several possible reasons for finding such a difference in “meaning”: the items really could be understood differently by subgroups, or the items could be tapping into a different latent trait in different groups, or these groups could affirm items at a different rate because of some other factor not accounted for by the model.

The US HFSSM was tested for DIF, and yet, whether or not the items were actually biased is controversial; Ohls et al. (25) reported that items were answered differentially by subgroups of the U.S. population, defined by race or ethnicity, region, and household composition, but concluded that the DIF was not great enough to affect the classification of a household's food security status. Wilde's (24) recent article on the matter argues otherwise, suggesting that households with and without children respond differently to certain questions in the measure, violating a key Rasch assumption and influencing how households are ultimately categorized. The controversy underscores the need to perform this type of testing where the intention of the scale is to aggregate responses across heterogeneous segments of the population. Because the FIMV Study was designed to develop scale items that could be applicable across 3 broad regions of Bangladesh, a DIF analysis was conducted to investigate the equivalence of items across 5 key subgroups, defined by land holdings (less than 0.5 acres or not), female literacy (female respondent read the literacy test or not), religion (Muslim or not), household demographics (family has children or not), and region. The DIF procedure can be performed only on items that have been shown to fit the model specifications for the population, so this analysis was conducted using the set of 20 items that remained following the assessment of item fit.

The next step in the procedure was to determine how the scale should be normalized using an internally consistent convention to

define the 0 point (24). Clearly, it is 1 thing to measure each item's severity by comparison to the severity of the item "*not eat bhalo mondo*" and another thing to measure each item's severity by comparison to the severity of "*had to eat wheat*." Given calibrations for 2 subgroups (say, landless and landed households) that exhibit some degree of DIF, the actual item or items that will be flagged depend on this choice of normalization. In some cases, one may be assured that most of the items measure the same underlying food security phenomenon but concerned that a small number of items may exhibit DIF. In such cases, the conventional practice in Rasch analysis is to choose a calibration that avoids unnecessarily discarding too many items. In the Bangladesh analysis, we followed this convention and chose an item with an intermediate level of severity as the reference item: the **t-th** item, "the main working adult skipped meals."

Item severity calibrations and their standard errors and the correlation of error residuals were determined for each population subgroup using WINSTEPS®. Based on this information, the standard error of the difference between the calibration of the same item in different subgroups was calculated in a spreadsheet along with the *t*-statistic and associated *P*-value. Item differences were flagged at the 95% confidence level. Because the final objective was to determine the set of items that would have a similar meaning across all subgroups, any item that was flagged, regardless of whether it behaved differently in 1 or all groups, was removed. The test was rerun a second time to verify that all remaining items were free of DIF across all subgroups.

Comparison of approaches

Following the completion of the separate qualitative and quantitative processes of item assessment, 2 lists were compiled of the items that were flagged for exclusion through each process. These lists were then compared to determine the extent to which the same items had been detected by the 2 different approaches. Next, the 2 lists of items that remained eligible for the scale were also examined to determine the extent to which they exhibited, and agreed on, Messick's (19) important elements of content validity: relevance and representativity (of the 4 food insecurity domains along a continuum of severity), technical quality, and similarity of meaning across the population. The level of agreement in the 2 scales was also assessed quantitatively by correlating their raw scores and by cross-tabulating the raw score terciles, using Pearson chi-square to assess the significance of the concordance.

RESULTS

Qualitative item assessment

Based on the combined inputs of the content experts and the respondent cognitive debriefings, 25 items were flagged for exclusion from the final scale. In addition to the evaluation criteria that had been defined for the content experts in advance, further criteria emerged in the course of the deliberations that were considered useful for detecting inappropriate items. **Table 2** presents those items that were excluded through the qualitative assessment process and the primary rationale for their exclusion.

Not universally applicable. The scale developers desired that all households be able to respond to all questions in the scale so that their response sets could be comparable. Therefore, all child-referenced items were left out of the overall scale because they were applicable to only two-thirds of the sample. Similarly, 2 items that asked about traditional Islamic practices of assistance to the food insecure—having to seek meat during the *Kurbani Eid* and taking charity (*jakat* or *fitra*)—were excluded because they were not relevant to Hindus or other non-Muslims.

Nonspecific to food insecurity. Items were excluded because they were characteristic of, but not specific to, the food insecure. For example, an item about consumption of "broken rice" was initially identified as a useful indicator of the "in-

adequate quality" aspect of household food insecurity (because it is an inferior form of rice). However, this item was debated when it became clear that most households, not just the food insecure, separate broken rice from a bag of good rice and cook it separately. In this case, consumption of a dish of broken rice would not in itself distinguish food-secure households from food-insecure ones.

Uncharacteristic of food insecurity. Content experts and respondents flagged 1 item, about not being able to afford snacks for the household, because it represented a behavior or perception that was not a characteristic of household food insecurity in Bangladesh. This question was excluded after it was agreed that even food-insecure households often managed a *taka* or so to buy their families an occasional snack. Because items were generated using in-depth ethnographic work that looked expressly for indications of food insecurity, this problem was limited to 1 item.

Unavailable and inaccessible: the "U-shaped curve." The feedback of content experts and respondents highlighted an important issue with the class of items dealing with socially unacceptable strategies to augment the household food supply (also known as "coping strategies"). Certain strategies may be exhausted, unavailable, or inaccessible to different segments of the population when the survey is administered, thus eliciting a response that does not reflect food insecurity but rather some other constraint. For instance, an item asking about selling or mortgaging productive assets would have a nonlinear relation to food security status; highly food-insecure households report no such experience because, despite their need, they have nothing of value to sell in exchange for food or cash. Food-secure households also respond negatively because they have no need to sell off assets in the first place. Borrowing money at high interest rates from a moneylender would have a similar relation to food security status for the same reasons. In referring to the most food-secure households, respondents reported, "they do not borrow money; rather they give loans." About the most food-insecure households, respondents confirmed that "they never take a loan because nobody wants to lend them money in fear they won't pay it back." Thus, only the households in the middle of the spectrum would be expected to affirm such an item.

Poor technical quality (meaning not understood as intended). Much of the intensive ethnographic work, pretesting, and pilot testing in the first phase of the study was aimed at ensuring that the meanings of questions were understood by respondents as intended by the researchers. Therefore, early on, initial item wording was revised frequently, double-barraged questions were eliminated, and response options were reworded. Still, the participants in the qualitative item assessment process found problems of a technical nature with a few of the questions, including the item "used money for another purpose to buy food," which attempted to convey the difficult concept of economic trade-offs, preferences, and sacrifice in the face of an overall budget constraint.

Differential meaning across population subgroups. Though the ethnographic research that informed the initial pool of candidate items was collected from 1 part of the country because of time and budget constraints, the respondent cognitive debriefings were from a geographically more varied sample, and the content experts had field experience from locations across the nation. Therefore, discussions with them uncovered potential geographically dependent differences in the meaning of items in different locations. For instance, *bhatar mar* (rice starch—the liquid left over after cooking rice) was initially thought to represent a food consumed rarely, by very hungry people. The same applied to *mishiti aloo* (sweet potato), *gom baja* (fried wheat), and *bon kochu* (wild taro). But then it was suggested that

TABLE 2

Items excluded from qualitatively derived scale and the basis for exclusion

| | Item description | Explanation for exclusion based on qualitative data | Exclusion category |
|-------|----------------------------|---|--|
| 1 | Sought Kurbanī meat | Not applicable to non-Muslims | Not universally applicable |
| 2 | Took charity | Not applicable to non-Muslims | Not universally applicable |
| 3 | Ate broken rice | Both food secure and insecure eat broken rice | Nonspecific to food insecurity |
| 4 | No hh snacks | Food insecure are typically able to spend a small amount of money for snacks | Uncharacteristic of food insecurity |
| 5 | No kid snacks | “Child” was not defined in the question, causing confusion and nonresponse. Not all families had children. | Poor technical quality/ not universally applicable |
| 6 | No big fish | Food secure households also face big fish supply constraint if not near market or pond. | Unavailable/ inaccessible – the “U-shaped curve.” |
| 7 | Ate water lily | A type of food eaten only in very severe situations – more severe than current sample | Different meaning across population subgroups |
| 8 | Sold items | Severely food insecure households have few or no items to sell or mortgage. | Unavailable/ inaccessible – the “U-shaped curve.” |
| 9 | Took money loan | Severely food insecure households not able to obtain loan | Unavailable/ inaccessible – the “U-shaped curve.” |
| 10 | Forego nonfood | Confusion over question wording | Poor technical quality |
| 11 | Ate sweet potato | Sweet potato is not universally indicative of food insecurity. Some households eat it because they like the taste. | Different meaning across population subgroups |
| 12 | Ate fried wheat | Fried wheat is not universally indicative of food insecurity across Bangladesh. Some households eat it because they like the taste. | Different meaning across population subgroups |
| 13 | Ate wild taro | Wild taro is not universally indicative of food insecurity across Bangladesh. | Different meaning across population subgroups |
| 14 | Ate rice starch | Rice starch is not universally indicative of food insecurity across Bangladesh. | Different meaning across population subgroups |
| 15 | No meat | Eating meat and eating “bhalo mondo” (rich food) are similar items. Meat is often considered “bhalo mondo.” | Poor technical quality |
| 16–24 | Child-referenced questions | About one-third of the sample did not have children under 5. | Not universally applicable |
| 25 | Ate wheat gruel | <i>Ata gola pani</i> is not universally indicative of food insecurity. Some households eat it because they like the taste. | Different meaning across population subgroups |

Source: Tufts FSNSP/FANTA data (2001).

individuals in some regions eat these foods not because of scarcity but because they like the taste. Nearly every potential example of a specific “less desired” or inferior food seemed to be a chosen (desired) food by someone somewhere else. An item about having to eat regular wheat (where rice is typically the preferred grain) faced a similar problem: food-secure families in some locations felt it was culturally desirable to eat wheat bread for at least 1 meal, and, because wheat in some locations was less expensive than rice, it was economically desirable too.

Rasch item assessment: model fit

Table 3 illustrates key characteristics of the 600-household survey sample. Table 4 presents the results of applying the Rasch model to the candidate items to determine the extent of model misfit. The item severity calibration column presents severity calibrations and their standard errors expressed in logits; they can be interpreted as the probability of an item being affirmed relative to the reference item, “working adult skipped meals.”

The infit and outfit columns display statistics used to detect model fit. Based on the criterion used in the development of the U.S. measure, which specified that any item exceeding an infit

and outfit of 1.2 should be modified or dropped from the scale, this analysis flagged only 2 items. The first, “give kids money for *shaidpati* (snacks),” had an infit/outfit of 1.64/2.81. The second, “could not purchase *chanachur* (a type of savory snack) and other snacks for the family,” had an infit/outfit of 1.4/9.73. After these 2 items had been dropped from the scale, and the model had been rerun, 1 additional item, “could not eat big fish,” no longer met the fit criteria. Whereas in the first iteration it had an infit/outfit of 1.07/2.54, the item had an infit/outfit of 1.28/8.65 after the second iteration. Once the “big fish” item was removed from the scale, all remaining items had infits/outfits well within the 1.2 range.

Rasch item assessment: differential item functioning

Table 5 summarizes the results of the DIF procedure as tested on key subpopulations. The first column presents the item severity calibrations for all households in the sample, relative to the reference item (working adult skips meals). The next columns disaggregate the item severity calibrations by subgroup and display the difference between the parameters for each subgroup household type.

TABLE 3

Key characteristics of survey sample

| Characteristic | N | Value* |
|--|-----|--------------------|
| Average household size | 600 | 5.3 (\pm 2.4) |
| Dependency ratio** | 600 | 1.9 (\pm 1.4) |
| Average age household head | 587 | 43.9 (\pm 13.9) |
| Female household headed (%) | 24 | 4.1 |
| Income sources per HH | 600 | 8.2 (\pm 3.3) |
| Household head literate (%) | 242 | 42.3 |
| Household head numerate (%) | 534 | 89.0 |
| Religion (%) | | |
| Muslim | 494 | 82.3 |
| Hindu | 105 | 17.5 |
| Other | 1 | 0.2 |
| Urban | 40 | 6.7 |
| Rural | 560 | 93.3 |
| Landless (< 0.5 acre) | 296 | 49.3 |
| WFP food insecurity ranking of sample upazila: | | |
| Very high food insecurity | 201 | 33.5 |
| High food insecurity | 120 | 20.0 |
| Moderate food insecurity | 259 | 43.2 |
| Low food insecurity | 20 | 3.3 |

Source: Tufts FSNP/FANTA data (2001).

* Values are percentages or means (\pm SEM).

** Dependency ratio is calculated as number of non-income earners/number of income earners.

None of the item calibrations differed significantly for households with and without children. Similarly, item calibrations did not depend on whether the female respondent in the household was literate or not. Because the literacy results were similar to the child household results, they are not presented in

Table 5. The severity calibrations of 4 items were significantly different for landless and landed households at the 0.05 level. These items asked about the inability to eat meat and *bhalo mondo* (good food), having to eat *mishti aloo* (sweet potato), and eating *ghom baja* (fried wheat). The calibrations for 2 of these items, referring to meat and *bhalo mondo*, were also significantly different across Muslim and non-Muslim households, along with the calibration for the item "borrowed staple grains from a shop."

Table 6 represents the results of same procedure applied to major geographic subgroups in the Northern, Central, and Southern regions of the country. Of the 3 regional comparisons, the Central versus the Northern region showed the fewest significant differences in item function: the calibrations of only 2 items, *personally skipped meals* and *worried about food*, were found to differ significantly. The regional pair with the most significant differences in item functioning was the South versus the North. Five item calibrations, including several of those that had been already flagged with DIF in other comparisons, appeared to function differently depending on the part of the country in which they were asked. Finally, the severity calibrations of 4 of the items administered in the Southern and Central regions also differed significantly, and each of these 4 had been flagged as having DIF in other subgroup comparisons. Once all of these items were excluded, the remaining items in the scale showed no additional evidence of DIF.

Comparison of qualitative and quantitative (Rasch) assessments

Twenty-three of the 38 candidate items were subjected to both the qualitative and Rasch assessments. Table 7 compares which of these 23 items were identified for exclusion from the

TABLE 4

Rasch item severity calibrations, infit and outfit statistics for food insecurity items

| Item | Response frequency (n = 581) | % | Item severity calibration | Real SE | Infit mean sq. | Outfit mean sq. |
|------------------------------|---------------------------------|-------|---------------------------|---------|----------------|-----------------|
| No rich food | 528 | 90.41 | -7.21 | 0.16 | 0.93 | 1.54 |
| No hh snacks | 462 | 79.11 | -5.99 | 0.12 | *1.4 | 9.73 |
| No meat | 440 | 75.34 | -5.68 | 0.12 | 0.77 | 1.43 |
| No big fish | 418 | 71.58 | -5.39 | 0.11 | **1.07 | 2.54 |
| Bought rice often | 241 | 41.27 | -3.33 | 0.11 | 0.93 | 1 |
| No kid snacks | 209 | 40.74 | -3.26 | 0.12 | *1.64 | 2.81 |
| Ate less food | 218 | 37.33 | -3.05 | 0.11 | 0.72 | 0.58 |
| Worried about food | 150 | 25.68 | -2.11 | 0.12 | 0.89 | 0.78 |
| Few square meals | 136 | 23.29 | -1.89 | 0.13 | 0.88 | 0.75 |
| Friends food loan | 122 | 20.89 | -1.65 | 0.13 | 0.83 | 0.58 |
| Skipped meals | 101 | 17.29 | -1.27 | 0.14 | 0.67 | 0.32 |
| Food ran out | 92 | 15.75 | -1.09 | 0.14 | 0.78 | 0.36 |
| Ate wheat | 79 | 13.53 | -0.81 | 0.15 | 1.07 | 0.75 |
| Took shop food loan | 78 | 13.36 | -0.78 | 0.15 | 1.08 | 1.25 |
| Working adult skipped meals | 49 | 8.39 | 0 | 0.18 | 0.87 | 0.33 |
| Ate rice starch | 37 | 6.34 | 0.43 | 0.2 | 0.82 | 0.27 |
| Ate broken rice | 35 | 5.99 | 0.52 | 0.2 | 0.98 | 0.38 |
| Ate wild taro | 12 | 2.05 | 1.95 | 0.32 | 0.86 | 0.35 |
| Ate fried wheat | 8 | 1.37 | 2.44 | 0.38 | 0.85 | 0.38 |
| Ate flour and water | 7 | 1.2 | 2.51 | 0.4 | 0.97 | 0.19 |
| Ate sweet potato | 6 | 1.03 | 2.77 | 0.44 | 1.12 | 0.38 |
| Food loan for guest/relative | 5 | 0.86 | 2.98 | 0.47 | 1.12 | 0.37 |
| Not eat for day | 4 | 0.68 | 3.23 | 0.52 | 1.11 | 0.35 |

Source: Tufts FSNP/FANTA data (2001).

* Infit and outfit > 1.2, removed from scale.

** Infit and outfit > 1.2 after removing both "could not purchase snacks" items, removed from scale.

TABLE 5

Rasch differential item functioning (DIF) of food insecurity items across key socio-demographic groups in Bangladesh

| Item name | Item severity calibrations (logits) | | | | | | | | | |
|------------------------------|-------------------------------------|-------------|--------|----------|----------|------------|--------|---------|--------|-------|
| | All | Non- Muslim | Muslim | DIF | Landless | Land-owner | DIF | No kids | Kids | DIF |
| No rich food | -10.11 | -13.17 | -9.64 | ** -3.53 | -7.93 | -13.47 | **5.54 | -9.7 | -10.24 | 0.54 |
| No meat | -7.47 | -10.17 | -7.06 | ** -3.11 | -6.6 | -9.12 | **2.52 | -7.15 | -7.57 | 0.42 |
| Bought rice often | -4.04 | -5.51 | -3.93 | -1.58 | -4.37 | -3.75 | -0.62 | -3.6 | -4.19 | 0.59 |
| Ate less food | -3.64 | -4.85 | -3.58 | -1.27 | -3.51 | -4.08 | 0.57 | -3.02 | -3.86 | 0.84 |
| Worried about food | -2.4 | -3.68 | -2.33 | -1.35 | -2.45 | -2.49 | 0.04 | -2.26 | -2.45 | 0.19 |
| Few square meals | -2.13 | -2.64 | -2.15 | -0.49 | -2.07 | -2.43 | 0.36 | -1.58 | -2.32 | 0.74 |
| Friends food loan | -1.85 | -3.37 | -1.74 | -1.63 | -1.82 | -2.07 | 0.25 | -1.58 | -1.95 | 0.37 |
| Skipped meals | -1.4 | -1.26 | -1.46 | 0.2 | -1.39 | -1.58 | 0.19 | -0.98 | -1.55 | 0.57 |
| Food ran out | -1.2 | -2.43 | -1.13 | -1.3 | -1.26 | -1.18 | -0.08 | -0.79 | -1.34 | 0.55 |
| Ate wheat | -0.88 | -1.63 | -0.86 | -0.77 | -0.77 | -1.27 | 0.5 | -0.59 | -0.98 | 0.39 |
| Took shop food loan | -0.85 | -3.2 | -0.62 | * -2.58 | -0.74 | -1.27 | 0.53 | -0.69 | -0.92 | 0.23 |
| Work adult skipped meals | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Ate rice starch | 0.47 | 1.25 | 0.45 | 0.8 | 0.34 | 1.41 | -1.07 | 0.82 | 0.35 | 0.47 |
| Ate broken rice | 0.56 | -1.63 | 0.74 | -2.37 | 0.79 | -0.16 | 0.95 | 1.28 | 0.35 | 0.93 |
| Ate wild taro | 2.07 | 0 | 2.21 | -2.21 | 2.29 | 1.41 | 0.88 | 2.42 | 1.96 | 0.46 |
| Ate fried wheat | 2.58 | 1.25 | 2.61 | -1.36 | 3.17 | 0.97 | *2.20 | 3.19 | 2.42 | 0.77 |
| Ate flour and water | 2.66 | 1.25 | 2.69 | -1.44 | 3.06 | 1.41 | 1.65 | 3.19 | 2.5 | 0.69 |
| Ate sweet potato | 2.92 | 0 | 3.17 | -3.17 | 3.77 | 0.97 | **2.80 | 4.46 | 2.61 | 1.85 |
| Food loan for guest/relative | 3.14 | 1.25 | 3.17 | -1.92 | 3.44 | 2.13 | 1.31 | 3.19 | 3.09 | 0.1 |
| Not eat for day | 3.39 | 1.25 | 3.43 | -2.18 | 3.44 | 3.35 | 0.09 | 2.42 | 3.87 | -1.45 |

Source: Tufts FSNSP/FANTA data (2001); * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

scale by each approach. The qualitative assessment of these 23 items flagged and discarded 10 items, leaving a 13-item scale. The Rasch assessment of infit and outfit flagged 3 items, and the Rasch test of differential item functioning flagged 10 items leaving a total of 10 items in the Rasch-derived scale. Only 7 of the same items, or 30% of the total, were excluded by both methods.

Table 8 compares the 13-item scale that was developed through the qualitative item assessment and the 10-item scale developed through the Rasch approach. In evaluating the 2 scales according to Messick's (19) criteria for content validity,

the scales derived through different methods both achieved, and were in broad agreement on, Messick's first 2 criteria—that the scale items should be relevant to and representative of the construct. These 2 criteria relate to the set of items as a whole and are not dependent on perfect agreement of the exact items in each scale. Both scales retained items relevant to each of the 4 aspects of the multifaceted household food insecurity experience. Both scales also retained items representing a range of severity of the household food insecurity problem.

Table 8 also shows that 8 of the items are also common to both scales. However, the remaining items differ, which implies

TABLE 6

Rasch differential item functioning (DIF) of food insecurity items across major geographic regions of Bangladesh

| Item name | Item severity calibrations (logits) | | | | | | |
|------------------------------|-------------------------------------|-------|---------|-------|----------|---------|---------|
| | All | South | Central | North | S-C DIF | S-N DIF | C-N DIF |
| No rich food | -10.11 | -9.72 | -10.02 | -11.6 | 0.3 | *1.88 | -1.58 |
| No meat | -7.47 | -6.43 | -8.38 | -8.27 | **1.95 | **1.84 | 0.11 |
| Bought rice often | -4.04 | -3.91 | -3.92 | -4.89 | 0.01 | 0.98 | -0.97 |
| Ate less food | -3.64 | -4.01 | -3.52 | -3.97 | -0.49 | -0.04 | -0.45 |
| Worried about food | -2.4 | -1.13 | -2.59 | -4.02 | **1.46 | ***2.89 | * -1.43 |
| Few square meals | -2.13 | -1.68 | -1.95 | -3.2 | 0.27 | **1.52 | -1.25 |
| Friends food loan | -1.85 | -2.11 | -2.24 | -1.49 | 0.13 | -0.62 | 0.75 |
| Skipped meals | -1.4 | -1.79 | -0.43 | -1.87 | ** -1.36 | 0.08 | * -1.44 |
| Food ran out | -1.2 | -0.72 | -1.71 | -1.65 | 0.99 | 0.93 | 0.06 |
| Ate wheat | -0.88 | 0.53 | -1.71 | -2.02 | **2.24 | **2.55 | -0.31 |
| Took shop food loan | -0.85 | -1.07 | -0.79 | -0.76 | -0.28 | -0.31 | 0.03 |
| Work adult skipped meals | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Ate rice starch | 0.47 | 0.45 | 0.17 | 1.05 | 0.28 | -0.6 | 0.88 |
| Ate broken rice | 0.56 | 1.36 | 0.58 | -0.65 | 0.78 | **2.01 | -1.23 |
| Ate wild taro | 2.07 | 2.12 | 3.56 | 1.84 | -1.44 | 0.28 | -1.72 |
| Ate fried wheat | 2.58 | 3.08 | 3.56 | 1.38 | -0.48 | 1.7 | -2.18 |
| Ate flour and water | 2.66 | 2.94 | 3.56 | 1.84 | -0.62 | 1.1 | -1.72 |
| Ate sweet potato | 2.92 | 4.17 | 2.33 | 1.38 | 1.84 | *2.79 | -0.95 |
| Food loan for guest/relative | 3.14 | 3.36 | 2.33 | 3.83 | 1.03 | -0.47 | 1.5 |
| Not eat for day | 3.39 | 3.36 | 3.56 | 3.83 | -0.2 | -0.47 | 0.27 |

Source: Tufts FSNSP/FANTA data (2001); * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

TABLE 7

Comparison of items excluded based on qualitative and Rasch assessments

| Item | Qualitative | Rasch infit/outfit | Rasch DIF |
|-----------------------------------|----------------|--------------------|-----------|
| No meat | X ¹ | | X |
| No kid snacks | X | X | |
| No household snacks | X | X | |
| Ate sweet potato | X | | X |
| Ate rice starch | X | | |
| Ate wild taro | X | | |
| Ate fried wheat | X | | X |
| Ate Flour and water | X | | |
| Ate broken rice | X | | X |
| No big fish | X | X | |
| Ate wheat | | | X |
| No rich food | | | X |
| Few square meals | | | X |
| Ate less food | | | |
| Skipped meals | | | X |
| No eat for day | | | |
| Working adult skipped meals | | | |
| Food ran out | | | |
| Worried about food | | | X |
| Bought rice often | | | |
| Took shop food loan | | | X |
| Friends food loan | | | |
| Food loan for guests or relatives | | | |

Source: Tufts FSNSP/FANTA data (2001).

¹ X = excluded.

a divergence in meeting the 2 content validity criteria pertaining to specific item characteristics (technical quality and similarly understood across the population). Though both assessment procedures had hurdles in place to ensure that items were of sufficient technical quality and that they had the same meaning across the population, the 2 approaches diverged regarding exactly which items did or did not demonstrate these characteristics. For instance, according to the Rasch approach, 3 items

TABLE 8

Comparison of Qualitative and Rasch-Derived Food Insecurity Scales by Food Insecurity Domain

| Food insecurity domain | Qualitatively derived scale items | Rasch derived scale items |
|------------------------|---|--|
| Quality | Few square meals >Ate wheat No rich food | Ate wild taro Ate flour and water Ate rice starch |
| Quantity | Ate less food Skipped meals Working adult skipped meals | Ate less food Working adult skipped meals |
| Acceptability | No eat for day Bought rice often Food loan for guests or relatives Took friends food loan Took shop food loan | No eat for day Bought rice often Food loan for guests or relatives Took friends food loan |
| Uncertainty | Worried about food Food ran out | Food ran out |

Source: Tufts FSNSP/FANTA data (2001).

in the qualitatively derived scale—personally skipping meals, worrying about food, and borrowing food from a shop—were classified as having a “different meaning” depending on the geographic location of the household even though the ethnographic assessment concluded that these items would be similarly understood across populations. Meanwhile, the qualitative approach detected items, including eating *bon kochu* (wild taro), *gom baja* (fried wheat), and *ata gola pani* (flour and water), that ethnographic assessment predicted would not be understood the same way across subpopulations. And yet, according to the test of DIF in Rasch, this latter group of items did not differ significantly in “meaning” from 1 area to the next.

To what extent do these 2 scales overlap in their identification of a household’s food insecurity status? Table 9 presents the results of a cross-tabulation of the household food insecurity scores, by tercile, produced by each method. The degree of concordance in the 2 scales is 90%, meaning that 90% of households were classified into the same tercile by the 2 methods (chi-square 810.90, $P < 0.0001$). Pearson correlation coefficient of the 2 scales was highly significant ($r = 0.96$, $P < 0.001$). The strength of the correlation varied slightly when assessed by population subgroup but was never less than $r = 0.94$, ($P < 0.001$).

DISCUSSION

In a review of 277 instances of scale development and reporting across 75 peer-reviewed journals, Hinkin (31) found that only 17% of the scales attempted to establish content validity. He concluded that “the generation of items may be the most important part of developing sound measures,” and yet, “the manner in which researchers report the item generation process may do a disservice because of the omission of important information regarding the origin of measures used” (p. 968). This paper compared a qualitative and quantitative (Rasch) method for selecting appropriate items for a measure of household food insecurity. A complementary objective was to make the process explicit to highlight the potential challenges in integrating mixed-method results.

Because most mixed-method research debates have focused primarily on issues of study design and the choice of disciplinary paradigm, there is little by way of methodologic blueprint to guide the researcher in drawing inferences from mixed methods. Erzberger and Kelle (32) outline 3 possible outcomes of integrating, and drawing inferences from, qualitative and quantitative methods engaged to address the same research question. The first possible outcome is that both results can converge, which would lend credence to both the empirical application of the methods as well as to the results and inferences based on these results. Second, results may relate to different phenomena or different aspects of the same phenomenon and, if they are complementary, may supplement each other like pieces in a puzzle. Third, the results of the 2 methods may be contradictory or divergent, casting doubt on both methods and concepts. This third, very common, outcome should not be cause for despair, however. Rossman and Wilson (33) point out that “searching for areas of divergent findings may set up the dissonance, doubt, and ambiguity often associated with significant creative intellectual insights” (p. 633). Thus, rather than wishing away divergent results, the authors of this paper welcomed such findings to illustrate the potential conflicts that can occur, and new insights derived, when mixed-method approaches do not agree.

In the case of Bangladesh, the 2 different approaches to developing the content of a food insecurity scale converged on

TABLE 9

Concordance in household food insecurity score tertile by qualitatively derived and quantitatively derived item assessment methods

| | | Tertiles of Rasch derived scale, % (n) | | | |
|--|-------|--|------------|------------|--------------|
| | | 1 | 2 | 3 | Total |
| Tertiles of qualitative derived scale, % (n) | 1 | 45.1 (270) | 1.2 (7) | 0 (0) | 46.2 (277) |
| | 2 | 4.3 (26) | 14.7 (88) | 2.7 (16) | 21.7 (130) |
| | 3 | 0 (0) | 1.8 (11) | 30.2 (181) | 32.1 (192) |
| | Total | 49.4 (296) | 17.7 (106) | 32.9 (197) | 100.00 (600) |

Source: Tufts FSNSP/FANTA data (2001).

certain, but not all, aspects of content validity. The 2 methods agreed in ultimately retaining items representing each of the 4 domains that had been found to capture the range of the food insecurity experience in Bangladesh. In the development of the US HFSSM, this was 1 major area of divergence. Though U.S. ethnographic information had identified socially unacceptable behaviors as an important part of the experience, the U.S. Rasch analyses flagged and excluded all items related to socially unacceptable strategies to augment the resource base as not fitting a single statistical dimension. In the Bangladesh study, both the qualitative and Rasch approaches retained items related to the social acceptability domain. The 2 approaches converged in excluding 7 of the same items and retaining 8 common items. This agreement between methods at the item level represents a reasonable degree of convergence of the 2 approaches in their assessment of technical quality and item equivalence across the population.

Where the 2 approaches and the final item sets diverged, what are possible reasons for the incongruities? One possibility is that the methods themselves, though set up to accomplish the same objective using similar criteria for item assessment, have different strengths and weaknesses in evaluating these criteria that could contribute to divergent results. When there is divergence over issues of within-sample generalizability, as there was in this study when the Rasch DIF flagged items that the qualitative approach did not, one may be more inclined to trust the results of the representative, quantitative technique. However, in this case there are other mitigating factors that favor the qualitative approach. The content expert panel, though purposively chosen, was composed of individuals familiar with the food insecurity experience across Bangladesh. And the respondents that participated in the cognitive debriefing mirrored a sizable cross section of the demographic and geographic subgroups that were also compared in the Rasch DIF procedure. Also, certain items that the Rasch DIF analysis rejected, such as skipping meals, have been found through ethnographic research in many other cultures to be central manifestations of the food insecurity experience (34). Therefore, given the experts' diversity of experience and intimate knowledge of local conditions, the qualitative item assessment likely produced a reasonably accurate picture of which items would have different "meaning" across the population.

On the other hand, despite the fact that the Rasch model results were based on a random representative sample of households from across Bangladesh, there are very persuasive reasons to question the item-level results of the Rasch assessment where they diverge from the qualitative results. Recall that proponents of the Rasch model posit that it describes a statistical reality to which the data either do or do not fit. As part of the underlying assumptions necessary to create this reality, the model specifies that all households facing the same

food insecurity constraint respond using a similar set of strategies in a similar configuration. It is highly plausible, based on the grounded ethnographic evidence, that the single-parameter Rasch model describes a statistical reality that is too rigid in its prescription or is incomplete or not sufficiently reflective of the household food insecurity situation as it really is. It is possible, then, that such items as worrying about food or skipping meals really "mean" the same thing to different population subgroups and that people in different regions are more or less likely to skip meals for some valid reason related to their locale, such as the acceptability of a management strategy, the availability of information informing their response, the degree of perceived risk, and the human capacity of the household members to utilize strategies at their disposal. These very logical empirically observed locale-specific differences are not accommodated under the assumptions of the Rasch model, meaning that items may be "flagged" and discarded that might not otherwise be problematic under a different statistical approach. Because of similar criticisms raised in the U.S. context (24), the continued use of the Rasch model is under consideration by the National Academies Panel contracted to review the US HFSSM approach (23).

Can the comparison of methods and results in this study produce a recommendation regarding which approach should be used in future attempts to develop the content of a scale? Based on the functional equivalence of the 2 household food insecurity scales, an argument could be made to use the more familiar method better suited to the skills and resources at hand, which, for many practitioners, will be the qualitative approach. However, it is important to recognize that the qualitative approach used in this paper was not a single focus group but rather a rigorous application of multiple sources of qualitative data in different geographic regions over several years. That said, in this study, the qualitative methods ultimately produced intuitively plausible results where the Rasch approach did not. The fact that these grounded insights conflict with the Rasch approach cannot discount the usefulness of Rasch altogether but can certainly cast doubt on the appropriateness of the Rasch model for the assessment of household food insecurity, suggesting that other, more flexible statistical models should be tested for developing household food insecurity scales. Given these interesting conclusions, the authors advocate for continuing to apply both qualitative and quantitative methods, in a systematic and transparent way, to lend additional credence to the results when they converge and to pinpoint the direction for further exploration where they do not.

ACKNOWLEDGMENTS

The authors would like to thank the Directors of Data Analysis and Technical Assistance, Inc., Md. Zahidul Hassan and Md. Zobair

and their staff for their incredible technical and logistical contributions; staff at World Vision/Bangladesh for their collaboration in data collection and testing; Mark Nord at USDA/ERS for useful discussion about the Rasch model; and Anne Swindale at AED/FANTA, Edward Frongillo at Cornell University, Nadra Franklin of AED, and Eunyoung Chung of USAID for their helpful comments on an earlier draft of this paper. Any mistakes are solely those of the authors.

LITERATURE CITED

1. Pelto GH, Freake HC. Social research in an integrated science of nutrition: Future directions. *J Nutr.* 2003;133:1231-4.
2. Hentschel J. Integrating the qual and the quan: when and why? In: Kanbur R, Ed. *Qualitative and quantitative poverty appraisal: complementaries, tensions, and the way forward.* Q-squared working paper 1. Ithaca, NY. 2005.
3. Brewer J, Hunter A. *Foundations of multimethod research: Synthesizing styles.* Thousand Oaks, CA: Sage Publications, Inc.; 2005.
4. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull.* 1959;56:81-105.
5. Radimer KL, Olson CM, Campbell CC. Development of indicators to assess hunger. *J Nutr.* 1990;120:1544-8.
6. Radimer KL, Olson CM, Greene JC, Campbell CC, Habicht JP. Understanding hunger and developing indicators to assess it in women and children. *J Nutr Educ.* 1992;24:S36-44.
7. Wolfe WS, Olson CM, Kendall A, Frongillo EA. Understanding food insecurity in the elderly: A conceptual framework. *J Nutr Educ.* 1996;28:92-100.
8. Wolfe WS, Olson CM, Kendall A, Frongillo EA. Hunger and food insecurity in the elderly—its nature and measurement. *J Aging Health.* 1998;10:327-50.
9. Wolfe WS, Frongillo EA, Valois P. Understanding the experience of food insecurity by elders suggests ways to improve its measurement. *J Nutr.* 2003;133:2762-9.
10. Quandt SA, Rao P. Hunger and food security among older adults in a rural community. *Hum Organ.* 1999;58:28-35.
11. Quandt SA, Arcury TA, McDonald J, Bell RA, Vitolins MZ. Meaning and management of food security among rural elders. *J Appl Gerontol.* 2001;20:356-76.
12. Melgar-Quinonez H, Kaiser LL, Martin AC, Metz D, Olivares A. Food insecurity among Californian Latinos: Focus-group observations. *Salud Publica Mex.* 2003;45:198-205.
13. Harrison GG, Stormer A, Herman DR, Winham DM. Development of a Spanish-language version of the U.S. Household Food Security Survey Module. *J Nutr.* 2003;133:1192-7.
14. Hamilton W, Cook J, Thompson W, Buron L, Frongillo E, Olson C, Wehler C. Household food security in the United States in 1995: Technical report of the Food Security Measurement Project. Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service; 1997.
15. Alaimo K, Froelich A. Measuring food insecurity and hunger, background paper for 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, National Research Council of the National Academies. Washington, DC: National Academies Press; 2004.
16. Coates J, Webb P, Houser R. Measuring food insecurity: Going beyond indicators of income and anthropometry. Washington, D.C: Food and Nutrition Technical Assistance Project, Academy for Educational Development; 2003.
17. Messick S. Validity. In: Linn R, editor. *Educational measurement.* 3rd ed. New York: American Council on Education, Macmillan Publishing Company; 1989.
18. Rubio DM, Berg-Weger M, Tebb SS, Lee ES, Rauch S. Objectifying content validity: Conducting a content validity study in social work research. *Soc Work Res.* 2003;27:94-105.
19. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol.* 1995;50:741-50.
20. Vogt DS, King DW, King LA. Focus groups in psychological assessment: Enhancing content validity by consulting members of the target population. *Psychol Assess.* 2004;16:231-43.
21. Someren MvBarnard Y, Sandberg J. *The think aloud method: A practical guide to modeling cognitive processes.* London: Academic Press; 1994.
22. Alaimo K, Olson CM, Frongillo EA. Importance of cognitive testing for survey items: An example from food security questionnaires. *J Nutr Educ.* 1999;31:269-75.
23. Committee on National Statistics of the National Academies, Panel to Review US Department of Agriculture's Measurement of Food Insecurity and Hunger. *Measuring food insecurity and hunger: Phase 1 report.* Washington, D.C.: National Academies Press; 2005.
24. Wilde PE. Differential response patterns affect food security prevalence estimates for households with and without children. *J Nutr.* 2004;134:1910-5.
25. Ohls J, Radbill L, Schirm A. *Household food security in the United States, 1995-1997: Technical issues and statistical report.* Princeton: Mathematica Policy Research, Inc.; 2001.
26. Bickel G, Nord M, Price C, Hamilton WL, Cook JT. *Guide to measuring household food security, revised 2000.* Alexandria, VA: Office of Analysis, Nutrition, and Evaluation, Food and Nutrition Service, U.S. Department of Agriculture; 2000.
27. Nord M, Satpathy AK, Raj N, Webb P, Houser R. Comparing household survey-based measures of food insecurity across countries: Case studies in India, Uganda, and Bangladesh. Boston: Tufts University Friedman School of Nutrition Science and Policy; 2002.
28. Smith EV Jr. Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *J Appl Meas.* 2001;2:281-311.
29. Bond T, Fox C. *Applying the Rasch model: Fundamental measures in the human sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.; 2001.
30. Linacre JM, Wright BD. Reasonable mean-square fit values. *Rasch Meas Trans.* 1994;8:370.
31. Hinkin TR. A review of scale development practices in the study of organizations. *J Management.* 1995;21:967-89.
32. Erzberger C, Kelle U. Making inferences in mixed methods: The rules of integration. In: Tashakkori A, Teddlie C, editors. *Handbook of mixed methods in social and behavioral research.* Thousand Oaks: Sage Publications; 2003.
33. Rossman GB, Wilson BL. Numbers and words: Combining quantitative and qualitative methods in a single-scale evaluation study. *Eval Rev.* 1985;5:627-43.
34. Coates J, Frongillo EA, Rogers BL, Webb P, Wilde PE, Houser RF. Commonalities in the experience of household food insecurity across cultures: What are measures missing? *J Nutr.* 2006;136:1438S-48S.