Published in final edited form as:

Nurs Res. 2013 ; 62(3): 149–159. doi:10.1097/NNR.0b013e31828eafe6.

Using the Rasch Measurement Model in Psychometric Analysis of the Family Effectiveness Measure

Linda L. McCreary, PhD, RN [Research Assistant Professor],
University of Illinois at Chicago College of Nursing, Chicago, Illinois

Karen M. Conrad, PhD, MPH, RN [Senior Research Scientist],
Program Metrics, LLC and Research Associate Professor, University of Illinois at Chicago School of Public Health, Oak Park, Illinois

Kendon J. Conrad, PhD, MSPH [Senior Research Scientist],
Chestnut Health Systems Lighthouse Institute and Professor Emeritus, University of Illinois at Chicago School of Public Health, Normal, Illinois

Christy K Scott, PhD [Regional Director],
Lighthouse Institute, Inc., a Division of Chestnut Health Systems, Chicago, Illinois

Rodney R. Funk, BS [Research Associate], and
Chestnut Health Systems Lighthouse Institute, Normal, Illinois

Michael L. Dennis, PhD
Senior Research Psychologist and GAIN Coordinating Center Director, Chestnut Health Systems, Inc., Normal, Illinois

Abstract

Background—Valid assessment of family functioning can play a vital role in optimizing client outcomes. Because family functioning is influenced by family structure, socioeconomic context, and culture, existing measures of family functioning--primarily developed with nuclear, middle class European American families--may not be valid assessments of families in diverse populations. The Family Effectiveness Measure was developed to address this limitation.

Objectives—To test the Family Effectiveness Measure with data from a primarily low-income African American convenience sample, using the Rasch measurement model.

Method—A sample of 607 adult women completed the measure. Rasch analysis was used to assess unidimensionality, response category functioning, item fit, person reliability, differential item functioning by race and parental status, and item hierarchy. Criterion-related validity was tested using correlations with five other variables related to family functioning.

Results—The Family Effectiveness Measure measures two separate constructs: The effective family functioning construct was a psychometrically sound measure of the target construct that was more efficient due to the deletion of 22 items. The ineffective family functioning construct consisted of 16 of those deleted items but was not as strong psychometrically. Items in both

Correspondence: Linda L. McCreary, PhD, RN, Research Assistant Professor, University of Illinois at Chicago College of Nursing (M/C 802), 845 S. Damen Avenue, Chicago, IL 60612-7350, Phone: (312) 355-3446, Fax: (312) 996-8945, mcreary@uic.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

constructs evidenced no differential item functioning by race. Criterion-related validity was supported for both.

Discussion—In contrast to the prevailing conceptualization that family functioning is a single construct, assessed by positively and negatively worded items, use of the Rasch analysis suggested the existence of two constructs. While the effective family functioning is a strong and efficient measure of family functioning, the ineffective family functioning will require additional item development and psychometric testing.

Whether the client is an individual member of a family or the client is the family as a whole, valid assessment of family functioning can play a vital role in optimizing client outcomes in nursing and other disciplines. Family functioning is influenced by many factors, including culture, socioeconomic status (SES), and family structure (McGoldrick & Hardy, 2008). The values that govern family interactions are rooted in cultural norms that influence the family's definition of effective family functioning (McGoldrick & Carter, 2003). Because cultural values may lead families to enact positive family functioning through ethnic-specific patterns of interactions, measures of family functioning need to allow assessment of those interactions.

Socioeconomic status exerts a widespread influence on family functioning through myriad effects on family composition, housing, educational and employment opportunities, and childrearing practices. (Hines, 1999). Family scientists need to take care when measuring family functioning to avoid confounding limited social and economic resources with ineffective family interactions. Family structure or composition vary widely and can influence who carries out the roles and activities of effective family functioning. Mother-headed single-parent families are increasingly prevalent in the United States, and frequently include both blood-related and fictive kin residing together or separately. Thus, measurement of family functioning should not be limited to interactions of co-residing blood or legal relatives but instead assess the interactions of all who are identified by the respondent as family members.

Although much is known about family functioning through measurement of nuclear families headed by married, middle-class, European American parents, relatively little empirical knowledge is available about family functioning within families of diverse cultural backgrounds, SES, and single-parent or extended-family structures. Moreover, evidence suggests that existing measures of family functioning, developed and validated primarily within a European American, middle-class cultural context (Bingenheimer, Raudenbush, Leventhal, & Brooks-Gunn, 2005) may not be a valid assessment of families in today's increasingly diverse population (Bingenheimer et al., 2005).

Many family functioning measures that are considered *approaching well-established* or *well-established* (Cohen et al., 2008) have demonstrated unacceptable psychometric properties when used with families from diverse cultures, low SES, or single-parent structures (Bingenheimer et al., 2005). For example, some items of the widely used Family Assessment Device (Epstein, Baldwin, & Bishop, 1983) confound low SES with ineffective family functioning (e.g., *We have trouble meeting our bills*; Tiffin, Pearce, Kaplan, Fundudis, & Parker, 2007), disproportionately lowering family functioning scores for all low-income families and failing to distinguish highly effective from ineffective low-income families. Further, the Family Environment Scale (Moos & Moos, 1976) has demonstrated unacceptably low internal consistency reliability with subscale α ranging .04–.34 (Friedemann, 1994) and .28–.62, and questionable factorial structure, particularly when used with low-income, African American people (Mazzeo, Mitchell, & Williams, 2008).

The role of family functioning as a correlate or predictor of many physical, mental, and behavioral health problems, including substance use disorders, clearly demonstrates the need for a valid measure (Wagner et al., 2010). To address the gap in measures valid for use with low-income African American families, the Family Effectiveness Measure (FEM) was developed (McCreary & Dancy, 2004; McCreary et al., 2008).

The purpose of this study was to evaluate the psychometric properties of the FEM using the Rasch measurement model supplemented by classical test approaches, and to make improvements based on results. The psychometric theories that guided this research were the Rasch measurement model and classical test theory. Specific aims of the study were to (a) assess unidimensionality, or whether the set of items represented a single construct; (b) examine response category functioning; (c) assess item fit; (d) assess person reliability; (e) assess differential item functioning (i.e., item invariance) across subgroups of women by race and parental status; (f) examine item hierarchy; and (g) examine criterion-related validity. The Rasch measurement model, discussed in detail below, is useful in the development of a new scale or revision of an existing scale. Complementing classical test theory-driven approaches, Rasch analysis offers a more powerful examination of item and scale performance than is available using classical test theory alone (Tennant & Conaghan, 2007).

Methods

Data Source

The FEM was administered as part of a battery of tests used in the parent study, a large intervention study of predominantly African American female offenders newly incarcerated for substance abuse violations (Scott & Dennis, 2012). Inclusion criteria for the parent study were: female, at least 18 years of age, with substance abuse or dependence in the 90 days before incarceration. Women were excluded if they lived or planned to move outside of Chicago, Illinois in the 12 months postrelease, were not fluent in English or Spanish, were released before the 14th day of incarceration, or were cognitively unable to provide informed consent.

Of the 810 women who met the parent study inclusion criteria, baseline data were analyzed from the 607 women who met an additional inclusion criterion: identifying at least one family member, defined as “anyone you consider as your family, regardless of a blood or legal connection.” Respondents completed the FEM, together with measures of the Global Appraisal of Individual Needs (GAIN) and several additional scales. The GAIN is a biopsychosocial assessment used in substance abuse treatment and research. All interviews were conducted by extensively trained, certified, and supervised interviewers. Extensive data quality checks were conducted. The de-identified dataset was downloaded from a secure password-protected web site and institutional review board approval was obtained from Chestnut Health Systems.

Sample Characteristics

The 607 respondents ranged in age from 18 to 63 years (mean = 37.2 ± 10.19). African American women comprised 82.9% ($n = 503$) of the sample; 7.7% ($n = 47$) were Caucasian, 4.8% ($n = 29$) were Hispanic, and 4.6% ($n = 28$) were mixed or other race. Opioids were the primary drug used by 45.1% ($n = 273$) of the sample, followed by cocaine (24.4%, $n = 148$), marijuana (16.7%, $n = 102$), alcohol (10.5%, $n = 64$), amphetamines (1.5%, $n = 9$) and other drugs (1.8%, $n = 11$). The majority (64.7%, $n = 393$) of respondents had at least one child under age 21; the mean number of children reported was 1.85 ± 2.01 (range 1 to 13). However, only 37.2% ($n = 226$) reported having been their child’s primary caregiver at the

time of incarceration. Over half (54.9%, $n = 333$) had less than a 12th grade education, 30.8% ($n = 187$) had 12th grade, and 14.3% ($n = 87$) had above 12th grade. Most (75.8%, $n = 457$) respondents had annual incomes 0–99% of the federal poverty threshold, 20.2% ($n = 122$) had incomes 100%–300% of the poverty threshold, and 4% ($n = 24$) reported higher incomes. Respondents were primarily never-married (71.4%, $n = 433$) or separated/widowed/divorced (16.2%, $n = 98$); 12.4% ($n = 75$) were in a couple relationship.

Measures

The original 42-item FEM was conceptualized as a unitary construct developed to assess positive family functioning, with the negative items reverse-scored to contribute to the assessment of positive family functioning. It was developed using qualitative research methods (40 interviews, 4 focus groups, and 35 cognitive interviews) with community-dwelling low-income African American single-parent families (McCreary & Dancy, 2004; McCreary et al., 2008). Initial item development and psychometric evaluation were conducted using classical test theory approaches (Nunnally & Bernstein, 1994). The 26 positively worded items (e.g., *Our family treats each other with respect*) were interspersed with the 16 negatively worded items (e.g., *Family members lie to each other*). Item responses *never true of us* to *always true of us* were on a 5-point Likert-type scale.

Using classical test theory (McCreary et al., 2008), the FEM demonstrated high internal consistency ($\alpha = .94$), good test-retest reliability ($r = .78$), and criterion-related validity with statistically significant ($p < .001$) moderate correlations in the expected directions with the Family APGAR (Smilkstein, 1978), the Psychological Aggression subscale of Conflict Tactics Scale-Parent-Child (Straus, Hamby, Finkelhor, Moore, & Runyan, 1998), and the Depression subscale of the Hospital Anxiety and Depression Scale (Zigmond & Snaith, 1983). Four positive (reciprocal material help when needed; a structured, safe environment for children; positive communication and problem solving; and a close-knit commitment to each other) and two negative (disrespectful and untrustworthy interactions; and a dangerous, neglectful environment for children) subscales were revealed within the construct using factor analysis (McCreary et al., 2008).

The criterion variables selected for the current study were based on those available in the data set that had strong theoretical connections and research supporting either positive or negative associations with family functioning. To provide evidence of the instrument's criterion-related validity, the FEM was correlated with measures of substance use problems, conflict tactics, depressive symptoms, satisfaction with family relationships, and social relationship quality of life. All of these scales except the quality of life subscale were part of the GAIN; reliability and validity evidence, adult norms and psychometric tables for GAIN scales are available (Dennis et al., 2008, pp. 1–8 to 1–12). All items except those of the quality of life scale were dichotomous (no/yes) variables.

The negative effects of substance use problems on family functioning have been documented (Leonard & Eiden, 2007; Townsend, Biegel, Ishler, Wieder, & Rini, 2006). Substance use was measured using the 16-item Substance Problem Scale (SPS), part of the GAIN, to assess problems related to substance use experienced by the participant in the past year. Cronbach's α reported by the developers of the GAIN for the SPS was .93. Cronbach's alpha for the SPS was .91 in this sample.

The use of aggressive and violent behaviors for handling interpersonal conflict has deleterious effects on family functioning (Katz & Woodin, 2002). This was assessed using the 12-item General Conflict Tactics Scale (GTCS) adapted for the GAIN from Straus (1990). Developers of the GAIN reported Cronbach's α for the GTCS was .86. In this study, Cronbach's alpha for the GTCS was .86.

Family functioning is impacted negatively by depression in a family member, particularly in a parent (Herr, Hammen, & Brennan, 2007). The 9-item Depressive Symptoms Scale (DSS), one subscale from the Internal Mental Distress Scale of the GAIN, was used to measure depression among respondents. Developers of the GAIN reported Cronbach's α of .89 for the DSS. Cronbach's α for the DSS was .89 with this sample.

Satisfaction with family relationships can be used to assess a family member's perception of family functioning (Smilkstein, 1978). A single item asking respondents if they were satisfied with their family relationships was used to operationalize this construct in this study.

Family functioning is related to quality of life (Greeff, 2000). Quality of life was assessed with the 3-item social relationships subscale of the World Health Organization Quality of Life (WHOQOL)-BREF, a shortened version of the WHOQOL-100 (Garcia-Rea & LePage, 2010). The developers of the WHO-BREF reported Cronbach's α ranging from .66 to .69 for the social relationships subscale. They recommended interpreting those values with caution, as they were based on three items, rather than the minimum of four items generally recommended for assessing internal consistency reliability. Cronbach's α for the social relationships subscale in the current study was .82.

The Rasch Measurement Model

Access to a new, larger data set presented the opportunity to conduct a more in-depth psychometric evaluation of the FEM using the Rasch measurement model. Rasch analysis is described in detail elsewhere (Bond & Fox, 2007; Hagquist, Bruce, & Gustavsson, 2009; Smith, Conrad, Chang, & Piazza, 2002). Briefly, it is the only item response theory model that has the desirable scaling properties of linear, interval measurement (Embretson & Reise, 2000). The Rasch model fulfills the requirements of fundamental measurement in that it generates a linear interval scale. Rasch analysis tests whether the data fit the model by assessing whether the response pattern observed in the data corresponds to the theoretical pattern expected by the model (Tennant & Conaghan, 2007). The Rasch model was selected because of the: (a) sound person-fit methods, (b) simplicity and ease of use since only one parameter is needed (important for wide clinical use and ease of explanation); (c) scaling properties of linear, interval measurement that enable a clear standard and simple graphic interpretation on person and item maps; (d) elimination of concerns about overfitting with additional parameters and perhaps missing important misfitting patterns (i.e., explaining away the misfit with added parameters); and (e) usefulness for smaller samples because it estimates only a single parameter (Conrad et al., 2012).

In Rasch analysis, the item hierarchy is created by the item difficulty estimates. Difficulty refers to the relative rarity with which an item is endorsed, with the more rarely endorsed items considered more difficult (Smith et al., 2002). In this analysis, the term *item endorsability* will be used in place of difficulty. The term *person ability* refers to the person's level of the construct being measured (e.g., level of family functioning). Item fit statistics are used to evaluate whether individual items contribute to unidimensionality. An examination of unidimensionality addresses the validity issue of construct-irrelevant variance. The model provides for an examination of the data (i.e., items and persons) for flaws or problems that are indicated by their failure to fit the model (Bond & Fox, 2007). As items and persons are placed on the same linear scale, the Rasch model provides a convenient framework for assessing scale invariance across subgroups through differential item functioning (DIF) analysis (Tennant & Conaghan, 2007).

The FEM was analyzed with the Rasch rating scale model (Rasch, 1980) to obtain linear interval measures. The Rasch rating scale model, a single parameter model used when the

construct to be measured is anticipated to be unidimensional, was selected. The Rasch rating scale model is appropriate when the data are polytomous, as from a Likert scale in which each item has the same number of response options and which applies one set of threshold values to all the items in the scale (Bond & Fox, 2007).

The rating scale model estimates the probability of person n choosing category level k on item i as a function of person n 's position on the latent trait, (β_n) and the endorsability of item i (δ_i) at the given threshold k (τ_k ; Bond & Fox, 2007). In the formula below, the probabilities are converted to log odds, where $\ln$ is the natural logarithm and the logit is the unit of measure (Bond & Fox, 2007, p. 282). Here, π_{nik} is the probability of respondent n endorsing item i at category level k , $1 - \pi_{nik}$ is the probability of respondent n not endorsing item i at category level k .

$$\ln \left[\frac{\pi_{nik}}{1 - \pi_{nik}} \right] = \beta_n - \delta_i - \tau_k$$

Analysis Plan

The analysis plan is organized around the specific aims. Winsteps Version 3.72 (<http://www.winsteps.com>) was used to perform the Rasch analysis and SPSS 19.0.0 was used to examine criterion-related validity.

Aim 1: Assess unidimensionality—Principal component analysis of residuals was used to examine whether a substantial factor existed in the residuals after the primary measurement dimension has been estimated (Smith, 2002). This allowed determination of whether the set of items represented a single construct. Analysis of the dimensionality was a two-step process. First, the measurement dimension of the scale was estimated using the Rasch model. The variance associated with this measurement dimension was extracted from the item-response data by computing standardized residuals: (observed - expected)/(model standard error). Second, a principal component analysis of the standardized residuals was used to determine whether substantial subdimensions existed within the items (Smith, 2002). If the items measure a single latent dimension as estimated by the Rasch model, then the remaining residual variance should reflect random variation.

Conservative criteria were used for judging unidimensionality. Although there are no hard rules for interpreting the results of principal component analysis of residuals, the first criterion for unidimensionality was that the variance explained by the measurement dimension be at least 40% (Linacre, 2006). The second criterion was that the variance explained by the first principal component of the residuals be no more than 15%. A minimum ratio of 3:1 was chosen for the variance in the measurement dimension compared to the variance of the first principal component of residuals according to Embretson and Reise (2000). Additional criteria were considered for unidimensionality using item fit statistics, as discussed below

Aim 2. Examine response category functioning—When using the rating scale model with items having more than two response options, such as a Likert-type format, it is important to examine the response category structure to evaluate how the response options are used (Tennant & Conaghan, 2007). When the responses to items correspond to the levels of the construct being measured, the thresholds between each response and the next are ordered in increasing value from low to high. If the thresholds do not progress in a linear fashion, or the distances between two response category thresholds are judged to be

inadequate, the category functioning may be improved by collapsing response categories (Tennant & Conaghan, 2007).

Aim 3. Assess item fit—Rasch analysis provides fit statistics to test assumptions of fundamental measurement (e.g., that high scorers should endorse or get right almost all of the easy items). Once identified, persons and items that *misfit* can then be examined qualitatively to determine the causes of the problems. Problems may include items with confusing wording or items that assess a construct that is different from the principal construct being measured. Understanding poor fit can guide decisions about improving or dropping items.

The Rasch model provides two indicators of misfit: *Infit* is sensitive to unexpected responses to items near the person's ability level and *outfit* considers differences between observed and expected responses regardless of how far away the item endorsability is from the person's ability (Tennant & Conaghan, 2007). Mean square fit statistics (MNSQ) are defined such that the model-specified uniform value of randomness is 1.0. Item fit indicates the extent to which the response to a particular item is consistent with the way the sample respondents have responded to the other items. In item fit analysis, MNSQ values between .75 and 1.33 are considered acceptable (Wilson, 2005). An item was considered misfit if both *infit* MNSQ and *outfit* MNSQ > 1.33.

Aim 4. Assess person reliability—The Rasch person reliability is based on the estimated locations of persons along the measurement continuum. While similar to Cronbach's alpha, person reliability is often lower because, unlike Cronbach's alpha, it does not include extreme scores (i.e., perfect scores) in the computation. The values are interpreted in the same manner. For both indices, the criterion for acceptability in a new measure was 0.75.

Aim 5. Assess differential item functioning for race and parental status—The Rasch model requires that subgroups having equal levels of the underlying construct should respond in a similar manner to the items measuring that construct (Tennant & Conaghan, 2007). Differential item functioning analysis allows examination of such item invariance. Items that demonstrate DIF should be investigated to determine what might be inferred about the underlying construct and what that implies about the samples of persons detected (Tennant & Conaghan, 2007). A conservative criterion of .5 logit difference was used in judging a meaningful DIF contrast (Wang, 2008). If the items are not invariant across subgroups, then changes to the measures can be considered (e.g., dropping items, developing new ones, or developing separate measures for specific subgroups).

Aim 6. Examine the item hierarchy—In Rasch analysis, the item hierarchy created by the item endorsability estimates provides an indication of construct validity; that is, the items are ordered in terms of endorsability (Smith et al., 2002). Rasch analysis allows persons and items, measured on the same logit scale, to be displayed simultaneously on a Wright map. The original FEM was developed using classical test theory to sample positive and negative aspects from the domain of family functioning, so there was no *a priori* hypothesis about the hierarchy of the items. The Rasch analysis in the current study provided an opportunity to examine the item hierarchy.

Aim 7. Examine the criterion-related construct validity—To examine criterion-related validity, a series of statistically significant zero-order correlations was hypothesized with five variables representing constructs theoretically related to family functioning

available in the data set. A Bonferroni multiple-significance-test correction of $p < .01$ was used.

Results

In assessing unidimensionality of the original 42-item FEM (Aim 1), the variance explained by the measure was 60.8%, and the variance explained by the first principal component of residuals was high (26.3%), forming a ratio of 2.31:1, which failed to meet the 3:1 criterion for unidimensionality. Examination of dimensionality revealed two clear dimensions that made sense theoretically, with 26 positively worded items in one group and 16 negatively worded (not reverse-scored) items in the other group. The FEM subsequently was analyzed as two scales, termed Effective Family Functioning (EFF) to assess the interactions associated with effective family functioning and Ineffective Family Functioning (IFF) to assess the interactions associated with ineffective family functioning. Study results are presented separately for the EFF and IFF. A weak, negative correlation between the two scales (Pearson's $r = -.218$) was further evidence of the presence of two dimensions (Kim Park, Garber, Ciesla, & Ellis, 2008).

For the EFF items, the variance explained by the measure was 60.4% and the variance explained by the first principal component of residuals was 12.2%, with a ratio of 4.95:1, exceeding all three of the criteria for unidimensionality. For the IFF, the variance explained by the measure was 42.1%, but the variance explained by the first principal component of residuals was 15.7%, just over the criterion of 15%; therefore the ratio was 2.68:1, slightly under the criterion of 3:1. Given that the IFF narrowly missed meeting the unidimensionality criteria, coupled with the realization that these results may have been influenced by the intermingling of EFF and IFF items and negative item wording, the decision was made to continue to analyze the 16-item IFF scale to inform future development of the IFF.

Next examined was the category functioning of the Likert 5-point response format with both the EFF and IFF (Aim 2). While there was the desired monotonic progression from one-step calibration to the next, the interval between step calibrations for the #2 and #3 response options (*half the time true of us* and *mostly true of us*) in both scales was much smaller than the recommended 1.4 to 5 logit interval (Linacre, 1999). Additionally, the curves lacked discrete peaks for each response option (Figure 1, curves A and C). Therefore, the #2 and #3 response options were combined, recoding to a 4-point response format. The intervals between the response categories were improved at approximately 3 logits, and the response categories conformed to the recommended Rasch Andrich thresholds. Each response threshold curve now had a discrete peak, indicating an area in which that response had the highest probability of being selected (Figure 1, curves B and D). Both the EFF and IFF showed clear step delineations with the 4-point response format.

Starting with the 26-item version of the EFF with the new 4-point response format, an iterative process was used whereby misfitting items were identified, examined, and deleted based on item analysis and expert judgment using the criteria of >1.33 for both infit and outfit (Aim 3). In the initial analysis of 26 items, four items exhibited misfit for both infit and outfit: Q13_ *Praise each other*, Q29_ *Depend on God*, Q37_ *Adults are in charge*, and Q40_ *Adults put kids first*. They were deleted after examining each item and judging that the content was not essential to the measure of the construct or was redundant with other items, examining interitem and item-total correlations, and verifying that coefficient alpha would not be decreased if the item was deleted. Then, analyses were rerun of the resulting 22-item version; one further misfitting item, Q19_ *Know what's going on with each other*, was deleted after item analysis. Analysis of the resulting 21-item version revealed one additional misfitting item, Q35_ *Teach kids how to be safe*, which was also deleted after item analysis.

The final 20-item EFF with the 4-point response format had no misfitting items and met the criteria for unidimensionality (Table 1).

Analysis of the 16-item IFF with the new 4-point response format indicated two misfitting items with both infit and outfit of >1.33 , Q7_If I need something, no one in my family will help and Q36_Leave the kids home alone. Upon examination, the content of these items was deemed essential to the measure of the construct and not redundant with other items, and there was concern that the negative wording of Q7 and the lack of age specificity of Q36 had created respondent confusion. Therefore, these items were not deleted, pending further item development and investigation (Table 2).

The Rasch person reliability for the EFF was .94; Cronbach's alpha was .97. For the IFF, the Rasch person reliability was moderately strong at .78; Cronbach's alpha was .85 (Aim 4).

For the EFF, the item hierarchy was invariant across all subgroups examined (Aim 5). There was no DIF for African Americans compared to all other races, or comparing respondents who were the parent of a child under age 21 years to those who were not parents. In examining DIF among the subgroups in the IFF, there was no DIF by racial group. However, one item exhibited DIF by parental status: Using the criterion of 0.5 logit difference, it was significantly easier for those with a child under 21 years of age to endorse Q7_If I need something, no one in my family will help (.23 logits vs. .82 logits).

The Wright maps for the two constructs, EFF and IFF, are depicted in Figure 2 (Aim 6). The left side of the map shows the distribution of respondents in the sample arranged along a hierarchy by person ability (i.e., their level of family effectiveness; Map A) or ineffectiveness (Map B). The right side of the map shows the array of items ordered by item endorsability. Persons at the top have the highest level of the construct being measured, and items at the top of the map are the most difficult to endorse (Lozano Rojas, Rojas Tejada, & Pérez Meléndez, 2009).

For both the EFF and IFF, the item hierarchy was very stable (data not shown). As noted above, there were no *a priori* hypotheses about the item hierarchy. For the EFF, the person mean was higher than the item mean, indicating that EFF items were relatively easy for this sample of substance-abusing women offenders to endorse. They found Q23_Trust each other and Q30_Spend time together among the most difficult and Q39_Teach kids what's important among the easiest to endorse. For the IFF, the person mean was lower than the item mean, indicating that the IFF items were relatively difficult for this sample to endorse. Item Q14_Scared a family member will hurt us was among the most difficult and Q27_Tell each other how to run their lives among the easiest to endorse.

Separate hypotheses were tested for the relationship between each construct and the five criterion-related variables (Aim 7; Table 3). As hypothesized, statistically significant negative correlations were found between the EFF and the Substance Problem Scale, General Conflict Tactics Scale, and the Depressive Symptoms Scale, and positive correlations were found with the Satisfaction with Family Relationships and the Social Relationships Quality of Life Subscale. Also as hypothesized, statistically significant positive correlations were found between the IFF and the General Conflict Tactics Scale, and the Depressive Symptoms Scale, and negative correlations were found with the Satisfaction with Family Relationships and all of the Quality of Life Subscales except the QOL-Psychological. Both the EFF and the original FEM, with reverse-scored negative items, were conceptualized to measure the construct of effective family functioning. The 20-item EFF and the original 42-item FEM had similar correlations with the criterion-related variables (Table 3).

Discussion

The Rasch analysis suggested that family functioning consists of two distinct constructs, effective functioning and ineffective functioning, requiring two separate measures. This finding is consistent with that of other researchers who have found that negatively worded items that were reverse-scored assessed a construct different from the one under study (Conrad et al., 2004; Rodebaugh, Woods, & Heimberg, 2007).

This finding is in contrast to the prevailing conceptualization that family functioning is represented by a single construct. Many measures of family functioning consist of positively and negatively worded items, with negative items reverse-scored and combined with positive items to yield an assessment of effective family functioning (e.g., the McMaster Family Assessment Device; Epstein et al., 1983). In other measures, all items are positive, representing effective family functioning, with low scores on items interpreted to represent ineffective family functioning (e.g., the Family APGAR; Smilkstein, 1978).

The existence of two separate constructs was further supported by the low negative correlation ($-.218$) between the EFF and the IFF. Families' scores on the EFF items may have little relationship to their scores on the IFF items. For example, family members may help each other out but may simultaneously leave the kids at home alone. Thus, a valid measure of ineffective family functioning cannot be obtained simply from low scores on a measure of effective family functioning. The results of this study contribute to the debate on how family functioning is conceptualized and measured in research and practice (Kim Park et al., 2008).

The psychometric properties of the EFF were very strong. The EFF evidenced clear unidimensionality. The decision to collapse the 5-category response structure to 4 categories was based on the small interval between step calibrations for the original #2 and #3 response options, and the lack of discrete peaks for the curve for those response options (Linacre, 1999). These findings suggested that the respondents were not distinguishing between those response options, so the two categories should be collapsed. The resulting 4-category response structure yielded a more meaningful sequence of categories in which each category probability curve in turn represents a distinct portion of the underlying construct (Bond & Fox, 2007).

Six misfitting EFF items were eliminated after the process of examining the item content and item statistics determined that the items were not essential to the measure of the construct. A clue to the reason for the misfit lay in the fact that persons with either low or high family functioning might endorse these items with equal ease (*Depends on God*) or difficulty (*Praise each other*). Even after reducing the number of items from 26 to 20, the internal consistency Rasch person reliability of the resulting 20-item EFF remained unchanged at .94. All 20 items were invariant across race and parental status, indicating that they function similarly across racial groups and for respondents who were and were not parents. Not surprisingly, given the toll that substance abuse takes on family relationships, this sample of substance abusing women offenders found Q23_ *Trust each other* and Q30_ *Spend time together* among the most difficult to endorse. Items reflecting the nurturing of children in the family were easiest to endorse, as could be expected; even as a parent is struggling with a substance abuse problem, family members tend to rally around her children to make sure their needs are met. The construct validity of the EFF was further supported by statistically significant correlations in the expected directions with all five criterion-related variables.

The psychometric results of the IFF were encouraging, but further work is needed. The negative items were interspersed with the positive items in the 42-item measure, which may

have influenced the psychometric results adversely. There was partial evidence of unidimensionality, and an examination of the item groupings revealed clustering around the relative seriousness of the ineffective family interactions, with those above the item mean on the Wright Map representing negative interactions of a more serious nature than those below the mean (Figure 2). The IFF showed acceptable person reliability. The 4-point response category functioned very well. The two misfitting IFF items, Q7_If I need something, no one in my family will help and Q36_Leave the kids home alone were retained pending item rewording; both represent essential content that taps the more difficult to endorse aspects of the construct, ineffective family functioning. The item-response wording may have contributed to the misfit by creating a double negative (Q7_If I need something, **no one in my family will help**, to which the response is **0 never** true of us). Rewording Q7 to *If I need something, my family will refuse to help* will eliminate navigating the double negative and revising Q36 to specify *children under age 12* may improve item clarity and performance. One of the two misfitting items (Q7_If I need something, no one in my family will help) also exhibited DIF by parental status. This item required navigating double-negative wording; rewording may improve its performance. All other IFF items were invariant. Not surprisingly, items that reveal the presence of such serious negative interactions as physical violence in the family (Q14_Scared a family member will hurt us), were among the most difficult to endorse and items that acknowledge the existence of petty or intrusive interactions, such as Q27_Tell each other how to run their lives, were among the easiest to endorse. Four of the five correlations between the criterion-related variables and the IFF supported construct validity, although the correlations were weaker than those between the EFF and the criterion variables.

The item hierarchies indicated that, as a group, the EFF items were relatively easy for these respondents to endorse and the IFF items were relatively difficult to endorse. Further qualitative research should be undertaken to better understand the logic of the item hierarchy.

Limitations and Strengths

As a self-report measure, there exists the potential for social desirability response bias. Respondents may feel the need to reflect an idealized positive image and minimize the negative image of their families. This may be an especially acute problem in a population of women offenders, who may feel marginalized and judged in a negative light. A further limitation was the administration of the negative items interspersed with the positive items, which was potentially confusing for respondents. This confusion might have attenuated the person reliability and criterion-related validity of the IFF. Like the original FEM evaluation study, the respondents in this study were predominantly African American. Although it was encouraging to find that the scales were invariant for non-African American respondents, the samples of non-African Americans were very small, and these results must be viewed as preliminary. A major strength of this study was the use of Rasch analysis, which allowed a critical psychometric analysis beyond that possible with classical test theory alone. A large data set with strict data quality control mechanisms in place was available. Finally, the criterion-related variables were well-established, psychometrically sound instruments.

Future Steps

Further testing of the EFF and the IFF should be done by administering them as separate measures, using the revised 4-category response format, and rewording items to eliminate navigating double negatives. Administered in this way, there should be even stronger psychometric results. Further qualitative work on the IFF is recommended to better understand this construct. In future work, the EFF and IFF scales should be further developed with larger samples drawn from other diverse populations. Assessing effective

and ineffective family functioning separately should allow nurses and others to better assess family strengths and weaknesses, and to support the evaluation of interventions designed to build upon and enhance family effectiveness while promoting strategies to lessen ineffective family interactions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Thank you to Barth R. Riley, PhD, for his thoughtful review of an earlier version of the paper. Funding was received for the parent research from the National Institutes of Health, National Institute of Drug Abuse; R01DA021174; Scott, PI.

References

- Bingenheimer JB, Raudenbush SW, Leventhal T, Brooks-Gunn J. Measurement equivalence and differential item functioning in family psychology. *Journal of Family Psychology*. 2005; 19:441–455. [PubMed: 16221024]
- Bond, TG.; Fox, CM. Applying the Rasch model: Fundamental measurement in the human sciences. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2007.
- Cohen LL, LaGreca AM, Blount RL, Kazak AE, Holmbeck GN, Lemanek KL. Introduction to special issue: Evidence-based assessment in pediatric psychology. *Journal of Pediatric Psychology*. 2008; 33:911–915. [PubMed: 16675715]
- Conrad KJ, Conrad KM, Mazza J, Riley BB, Stein MA, Funk R, Dennis ML. Dimensionality, hierarchical structure, age generalizability, and criterion validity of the GAIN's Behavioral Complexity Scale. *Psychological Assessment*. 2012
- Conrad KJ, Wright BD, McKnight P, McFall M, Fontana A, Rosenheck R. Comparing traditional and Rasch analyses of the Mississippi PTSD Scale: Revealing limitations of reverse-scored items. *Journal of Applied Measurement*. 2004; 5:1–16. [PubMed: 14757988]
- Dennis, ML.; White, M.; Titus, JC.; Unsicker, J. Global Appraisal of Individual Needs: Administration Guide for the GAIN and Related Measures (Version 5). Normal, IL: Chestnut Health Systems; 2008. Retrieved on August 30, 2012 from http://www.gaincc.org/_data/files/Instruments%20and%20Reports/Instruments%20Manuals/GAIN-I%20manual_combined_0512.pdf
- Embretson, SE.; Reise, SP. Item response theory for psychologists. Mahwah, NJ: Erlbaum; 2000.
- Epstein NB, Baldwin LM, Bishop DS. The McMaster Family Assessment Device. *Journal of Marital & Family Therapy*. 1983; 9:171–180.
- Friedemann ML. Evaluation of the Congruence Model with rehabilitating substance abusers. *International Journal of Nursing Studies*. 1994; 31:97–108. [PubMed: 8194940]
- Garcia-Rea EA, LePage JP. Reliability and validity of the World Health Organization Quality of Life: Brief Version (WHOQOL-BREF) in a homeless substance dependent veteran population. *Social Indicators Research*. 2010; 99:333–340.
- Greeff AP. Characteristics of families that function well. *Journal of Family Issues*. 2000; 21:948–962.
- Hagquist C, Bruce M, Gustavsson JP. Using the Rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies*. 2009; 46:380–393. [PubMed: 19059593]
- Herr NR, Hammen C, Brennan PA. Current and past depression as predictors of family functioning: A comparison of men and women in a community sample. *Journal of Family Psychology*. 2007; 21:694–702. [PubMed: 18179341]
- Hines, PM. The family life cycle of African American families living in poverty. In: Carter, B.; McGoldrick, M., editors. *The expanded family life cycle: Individual, family, and social perspectives*. 3rd ed. Boston: Allyn and Bacon; 1999. p. 327–345.

- Katz LF, Woodin EM. Hostility, hostile detachment, and conflict engagement in marriages: Effects on child and family functioning. *Child Development*. 2002; 73:636–651. [PubMed: 11949913]
- Kim Park IJ, Garber J, Ciesla JA, Ellis BJ. Convergence among multiple methods of measuring positivity and negativity in the family environment: Relation to depression in mothers and their children. *Journal of Family Psychology*. 2008; 22:123–134. [PubMed: 18266539]
- Leonard KE, Eiden RD. Marital and family processes in the context of alcohol use and alcohol disorders. *Annual Review of Clinical Psychology*. 2007; 3:285–310.
- Linacre JM. Investigating rating scale category utility. *Journal of Outcome Measurement*. 1999; 3:103–122. [PubMed: 10204322]
- Linacre JM. Data variance explained by measures. *Rasch Measurement Transactions*. 2006; 20:1045–1047.
- Lozano Rojas OM, Rojas Tejada AJ, Pérez Meléndez C. Development of a specific health-related quality of life test in drug abusers using the Rasch rating scale model. *European Addiction Research*. 2009; 15:63–70. [PubMed: 19142005]
- Mazzeo SE, Mitchell KS, Williams LJ. Anxiety, alexithymia, and depression as mediators of the association between childhood abuse and eating disordered behavior in African American and European American women. *Psychology of Women Quarterly*. 2008; 32:267–280.
- McCreary LL, Crittenden KS, Dancy BL, Ferrans CE, Fu Q, Smith EV. Psychometric evaluation of the Family Effectiveness Measure among low-income African American single-parent families. *Western Journal of Nursing Research*. 2008; 30:1021–1022.
- McCreary LL, Dancy BL. Dimensions of family functioning: Perspectives of low-income African American single-parent families. *Journal of Marriage and Family*. 2004; 66:690–701.
- McGoldrick, M.; Carter, B. The family life cycle. In: Walsh, F., editor. *Normal family processes: Growing diversity and complexity*. 3rd ed. New York: Guilford Press; 2003. p. 375–398.
- McGoldrick, M.; Hardy, KV. *Re-visioning family therapy: Race, culture, and gender in clinical practice*. 2nd ed. New York: Guilford Press; 2008.
- Moos RH, Moos BS. A typology of family social environments. *Family Process*. 1976; 15:357–372. [PubMed: 1026454]
- Nunnally, JC.; Bernstein, IH. *Psychometric theory*. 3rd ed. New York: McGraw-Hill; 1994.
- Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press; 1980.
- Rodebaugh TL, Woods CM, Heimberg RG. The reverse of social anxiety is not always the opposite: The reverse-scored items of the Social Interaction Anxiety Scale do not belong. *Behavior Therapy*. 2007; 38:192–206. [PubMed: 17499085]
- Scott CK, Dennis ML. The first 90 days following release from jail: Findings from the Recovery Management Checkups for Women Offenders (RMCWO) experiment. *Drug and Alcohol Dependence*. 2012; 125:110–118. [PubMed: 22542465]
- Smilkstein G. Family APGAR: A proposal for a family function test and its use by physicians. *Journal of Family Practice*. 1978; 6:1231–1239. [PubMed: 660126]
- Smith EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*. 2002; 3:205–231. [PubMed: 12011501]
- Smith EV, Conrad KM, Chang K, Piazza J. An introduction to Rasch measurement for scale development and person assessment. *Journal of Nursing Measurement*. 2002; 10:189–206. [PubMed: 12885145]
- Straus, MA. The Conflict Tactics Scale and its critics. In: Straus, MA.; Gelles, RJ., editors. *Physical violence in American families*. New Brunswick, NJ: Transaction Publishers; 1990. p. 49–73.
- Straus MA, Hamby SL, Finkelhor D, Moore DW, Runyan D. Identification of child maltreatment with the Parent-Child Conflict Tactic Scales: Development and psychometric data for a national sample of American parents. *Child Abuse & Neglect*. 1998; 22:249–270. [PubMed: 9589178]
- Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied and what one should look for in a Rasch paper. *Arthritis & Rheumatism*. 2007; 57:1358–1362. [PubMed: 18050173]

- Tiffin PA, Pearce M, Kaplan C, Fundudis T, Parker L. The impact of socio-economic status and mobility on perceived family functioning. *Journal of Family Economic Issues*. 2007; 28:653–667.
- Wagner KD, Ritt-Olson A, Chou CP, Pokhrel P, Duan L, Baezconde-Garbanati L, Unger JB. Associations between family structure, family functioning, and substance use among Hispanic/Latino adolescents. *Psychology of Addictive Behaviors*. 2010; 24:98–108. [PubMed: 20307116]
- Wang WC. Assessment of differential item functioning. *Journal of Applied Measurement*. 2008; 9:387–408. [PubMed: 19092232]
- Wilson, M. *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum Associates; 2005.
- Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*. 1983; 67:361–370. [PubMed: 6880820]

Figure 1.
Category Probability Curves. A. EFF 20 items, 5 response options; B. EFF 20 items, 4 response options; C. IFF 16 items, 5 response options; D. IFF 16 items, 4 response options

Figure 2.

Wright Maps simultaneously displaying item difficulty (endorsability) hierarchy and person ability (level of family functioning). Map A. Effective Family Functioning Scale: 20 Items with 4 Response Categories ($n = 581$); Map B. Ineffective Family Functioning Scale (IFF) 16 Items with 4 Response Categories ($n = 607$)

Table 1

Items, Item Measures, and Item Fit Statistics for the Effective Family Functioning Scale

Item Number with Label	Abbreviated Item Wording ^a	Rasch Measure (in logits)	Infit MNSQ ^b	Outfit MNSQ
Q1_HELP_EACH_OTHER	Help each other out	0.12	1.01	0.99
Q2_SHARE_THE_WORK	Share the work of keeping things going	0.30	0.88	0.88
Q3_GET_NEEDS_FAMILY	When need something, can get it from family	0.12	0.87	0.82
Q5_PROVIDE_FAMILY_NEEDS	Adults able to provide family's needs	0.13	0.93	0.94
Q6_DO_FOR_EACH_OTHER	Do things for each other	-0.05	0.87	0.86
Q8_SHOW_AFFECTION	Show each other affection	0.41	0.86	0.86
Q9_GET_ALONG	Get along with each other	0.64	1.10	1.12
Q11_COUNT_ON_FAMILY	When we have trouble, we can count on family	0.30	1.05	1.08
Q15_WILL_LISTEN	When we need to talk, someone will listen	0.06	0.84	0.80
Q16_KIDS_CAN_TALK	Kids feel they can talk to family adults	0.06	1.03	1.00
Q17_TRY_SOLVE	Try to solve our problems with each other	-0.02	0.82	0.77
Q21_FUN_TOGETHER	Have fun together	0.40	0.95	0.96
Q23_TRUST	Trust each other	0.64	1.07	1.06
Q25_STICK_TOGETHER	Stick together	-0.15	1.30	1.37
Q30_SPEND_TIME_TOGETHER	Spend time together	0.76	1.14	1.16
Q32_RESPECT	Respect each other	0.23	1.08	1.10
Q33_DONT_GIVE_UP_OTHER	Don't give up on each other	-0.66	1.07	1.01
Q34_MAKE_SAFE_PLACE	Make home safe for kids	-1.37	0.97	0.94
Q38_HAVE_STABILITY_LIVES	Have stability in our lives	-0.59	1.09	1.12
Q39_TEACH_WHAT_IMPT	Teach kids what's important	-1.34	0.94	0.85

Notes. Deleted from original 26-item Effective Family Functioning Scale due to misfit (data not shown): Praise each other; Know what's going on with each other; Depend on God; Teach kids how to be safe; Adults are in charge; and Adults put kids first

^a All items are protected by copyright and may be used with permission of first author. For permission and full item wording, please contact Linda McCreary: mcreary@uic.edu

^b MNSQ = Mean square fit statistics

Table 2

Items, Item Measures, and Item Fit Statistics for the Ineffective Family Functioning Scale

Item Number with Label	Abbreviated Item Wording ^a	Rasch Measure (in logits)	Infit MNSQ ^b	Outfit MNSQ
Q4_BREAK_PROMISE	Break our promises	-1.05	0.74	0.76
Q7_NO_FAMILY_HELP	No one in family will help	0.42	1.74	1.77
Q10_KIDS_FEEL_BAD	Adults make kids feel bad	0.86	1.41	1.23
Q12_FIND_FAULT	Find fault with each other	-1.01	0.69	0.68
Q14_SCARED_HURT_BY_FAMILY	Scared a family member will hurt us	1.75	1.50	1.15
Q18_STAY_ANGRY	Stay angry at each other	0.06	0.95	0.91
Q20_ARGUE	Argue with each other	-1.17	0.63	0.65
Q22_LIE_TO_OTHERS	Lie to each other	-0.94	0.65	0.66
Q24_TELL_SECRET	Tell others' secrets to outsiders	-1.05	1.08	1.09
Q26_FEEL_HOPELESS	Feel hopeless.	0.66	1.02	0.92
Q27_TELL_RUN_LIFE	Tell each other how to run their lives	-1.21	1.07	1.09
Q28_TELL_KIDS_OPPOSITE	Family members interfere with our parenting	-0.36	1.10	1.08
Q31_STEAL_FROM_EACH_OTHER	Steal from each other	0.53	1.22	1.06
Q36_KIDS_HOME_ALONE	Leave the kids home alone	1.07	1.76	1.56
Q41_KIDS_SHOULDER_TOO_MUCH	Kids have too much responsibility	0.96	1.17	0.92
Q42_KIDS_OUT_OF_CONTROL	Kids out of control	0.49	1.07	0.97

Notes. Ineffective Family Functioning items Q7 and Q36 misfit using the criterion of both infit MNSQ and outfit MNSQ >1.33.

^a All items are protected by copyright and may be used with permission of first author. For permission and full item wording, please contact Linda McCreary: mcreary@uic.edu

^b MNSQ = Mean square fit statistics

Table 3

Correlations among Effective Family Functioning, Ineffective Family Functioning, original Family Effectiveness Measure and Criterion-Related Validity Variables (n =607)

Criterion-Related Variables	EFF (20 items)	IFF (16 items)	Original FEM (42 items) ^a
	<i>r</i>	<i>r</i>	<i>r</i>
EFF	–		.896 [*]
IFF	–.218 [*]	–	–.519 [*]
SPS	–.130 [*]	.016	–.074
GCTS	–.139 [*]	.137 [*]	–.169 [*]
DSS	–.235 [*]	.122 [*]	–.227 [*]
Satisfaction with Family Relationships	.326 [*]	–.123 [*]	.318 [*]
QOL-Social	.339 [*]	–.144 [*]	.346 [*]

Notes. Shaded areas indicate nonsignificant correlations.

EFF = Effective Family Functioning, IFF = Ineffective Family Functioning, FEM = Original Family Effectiveness Measure, SPS = Substance Problem Scale, GCTS = General Conflict Tactic Scale, DSS = Depressive Symptom Scale, QOL = Quality of Life

^{*} Significant at $p < .01$ level (1-tailed).

^a The 16 items of the FEM used to assess ineffective family functioning were reverse-scored to a positive valence when included as part of the measure of effective family functioning; therefore, negative correlations of the FEM with the IFF, SPS, GCTS, and DSS are to be expected.