Applications of Item Response Theory (IRT) Modeling for Building and Evaluating Questionnaires Measuring Patient-Reported Outcomes

Bryce Reeve, Ph.D.
National Cancer Institute

IRT modeling provides questionnaire developers a powerful tool for evaluating the properties of their surveys both at the item and scale level, and allows them the ability to tailor an instrument for an individual (i.e., utilizing computerized-adaptive testing) or a study population. This summary paper introduces the various tools that are available for researchers to examine questionnaire properties and the following paper, written by Dr. Maria Orlando, will address the critical issues one must consider in fitting IRT models to health outcomes and behavior data including model assumptions, model fit, and sample size requirements. Both this and Dr. Orlando's papers are part of a series of papers on IRT modeling and each paper, currently, does not stand alone. Dr. Hambleton's paper discusses the concepts and special features of IRT modeling. Drs. Teresi, Morales, and Fleishman discuss the applications of IRT modeling for evaluating measurement equivalence when questionnaires are adapted to new languages or populations. Dr. Dorans paper talks about the utility of IRT modeling for linking or equating heath outcomes questionnaires. Finally, Drs. Chang, Bjorner, and Thissen discuss how IRT modeling is used to build and maintain item banks which serve as the foundation for computerized-adaptive testing.

Throughout this summary paper, examples of the IRT tools for examining item and scale properties will be illustrated from a study on breast cancer survivors' responses to Dr. Northouse's Fear of Recurrence (FOR) scale. The FOR scale consists of 22 items with a 5-point Likert-type scale: "Strongly Agree", "Agree", "Neutral", "Disagree", "Strongly Disagree". Items were scored so that higher numbers reflect increased cancer survivor's fear of disease recurrence. A number of IRT models exist for modeling polytomous response data (see Orlando's paper and the Table of IRT Models). After evaluating model assumptions such as unidimensionality, I choose to use Samejima's Graded Response Model and Masters' Partial Credit Model for illustrations used in this manuscript.

Evaluating Properties of Items in a Unidimensional Scale

Item Characteristic Curves

Item Characteristic Curves (ICC, also known as trace lines, category response curves, or probability curves) model, in probabilistic terms, the relationship between a person's response to a question and his or her level on the construct (symbolized by θ) being measured by the scale. This relationship is conditional in that people with higher levels on the underlying construct will have a higher probability of endorsing response categories that are consistent with higher trait levels. Figure 1 presents the ICCs, estimated using the IRT Graded Response Model, for the question, "I do not worry about my illness returning." Breast cancer survivors with low fear of disease recurrence had a high probability to endorse "strongly agree" and women with high fear of disease recurrence chose "strongly disagree".

ICC's provide a wealth of information about item properties. The height (or steepness) of the curves reflect the discrimination ability of the item. In other words, how well the item's five response categories discriminate among women with different levels of fear. The discrimination (or slope) parameter (*a*) is analogous to an item's correlation with the total score in classical test theory. Steeper slopes indicate that smaller changes along the construct continuum will reflect larger changes in item endorsement probabilities. The locations where response curves intersect along the construct continuum reflect an item's difficulty[1] or severity. ICC curve intersection points are determined by the IRT GRM's threshold (*b*) parameters. With the FOR scale, low threshold parameter estimates ($b < 0$) indicate that those response categories are likely to be endorsed by women with low fear of disease recurrence, and vice versa.

ICCs offer questionnaire developers the ability to evaluate the appropriateness of the response format used in the questionnaire. In Figure 1, the "Neutral" response option is overshadowed by its neighbor categories "agree" and "disagree" indicating that at no point along the construct continuum is a person likely to answer "neutral" over any other category. This finding suggests that the response option may be dropped in revised versions of the questionnaire. ICCs can also point out when more response categories are needed which is reflected when a response category is the dominant choice (highest likelihood) across a large portion of the underlying construct continuum (θ).

Samejima's Graded Response Model was used for the above illustrations which assumed a pre-specified order of response categories (i.e., from "Strongly Agree" to "Strongly Disagree"). However, if category order is not clear (sometimes the case with nominal-type categories), one can evaluate category order using Bock's Nominal Response Model, Andrich's Rating Scale Model, Masters' Partial Credit Model, or Muraki's Generalized Partial Credit Model.

Item Information Curves

Item information curves (or functions) indicate the range over θ where an item is best at discriminating among individuals. Higher information, determined by the discrimination parameter, denotes more precision (or reliability) for measuring a person's trait level. Figure 2 displays the information functions for three items in the FOR scale. The questions "I think more

[1] "Difficulty" is borrowed from educational assessment, where the goal is to match item difficulty with student ability. For example, easy math problems can be solved by students with low math ability and hard math problems can only be answered by students with high math abilities.

about my health now than before my illness was diagnosed" and "When I think about my future health status, I feel some uneasiness" provide more information over most fear levels than the question "I do not worry about my illness returning." The threshold parameters determine the location of the information along theta, with greater information located at the intersections between ICC curves. Of interest in Figure 2, the question "When I think about my future health status, I feel some uneasiness" (dashed curve) provides more information for measuring women with high fears as compared to the other two items.

Multiple item information curves can be overlaid and examined in a single graph which allows instrument developers to observe how items function in relation to each other. For future development of revised questionnaires, developers can pick and choose items based on the magnitude and location of information provided by the scale items. Items with low information or items that provide duplicate information may be dropped from the scale to create a shortened version of the instrument.

Item – Person Maps

An Item-Person Map[2] locates along the θ continuum where the sample respondent estimated levels (or scores) line up with the average difficulty (or location) of the items all on the same metric. Item-person maps are mostly produced by Rasch model software whose models constrain all items in the scale to have equal levels of discrimination, thus items can be compared with one or another in terms of difficulty or location without considering discrimination ability. Figure 3 presents the item-person map for the 22 items in the FOR scale. On the left side of the vertical dashed line, "X" represents the estimated level (score) of a person along the Fear continuum with women at the top of the figure scoring in the high fear range. On the right side

[2] Item - person maps have been relabeled Wright Maps in honor of Dr. Benjamin Wright of the University of Chicago Department of Education

of the vertical dashed line, average item difficulties are presented with more severe (difficult) items at the top (items are identified with labels from item 1, V1, to item 22, V22).

Item-person maps are great tools to see how well the scale items match up with the respondents being measured. One can evaluate the breadth of the scale by observing if the range of item difficulties covers the estimated women scores or identify potential gaps between items along the measured construct continuum. Figure 3 suggest that a few women with high fear at the top of the scale and low fear at the bottom of the scale are not measured reliably because of lack of item coverage. One can also use this map to select items that measure different ranges (or regions) of the continuum or to remove items where there is redundant content coverage.

Differential Item Functioning (DIF) Testing

DIF occurs whenever one group consistently responds differently to an item than another group. In other words, respondents with similar levels of θ have different probability of responding to an item according to their population membership. Scales containing such items have reduced validity for between-group comparisons because their scores are influenced by a variety of attributes other than those intended. IRT provides an attractive framework for DIF testing, which is discussed in detail in the summary papers written by Drs. Jeanne Teresi, Leo Morales, and John Fleishman.

Evaluating Properties of a Unidimensional Scale

Scale Information Curve

The scale (or test) information curve (or function) indicates the level of information (i.e., reliability) provided by the scale over the range of the construct continuum. The scale information curve is simply the sum of the item information curves. Figure 4 presents the scale information function for the 22-item FOR scale. Along the vertical axis, the associated level of reliability associated with different information magnitudes is presented. Overall, the FOR scale is very reliable for measuring most women's level of fear; however measurement precision drops off in the high fear range.

Standard Error of Measurement Curve

The Standard Error of Measurement (SEM) curve evaluates the precision of the scale to measure people at different levels along the construct continuum. The standard error of measurement is simply defined as $\mathrm{SEM} = 1 \big/ \sqrt{information}$, and varies conditional on θ. Thus, the SEM curve provides similar detail as the scale information curve, but expressed in terms of standard error.

Revising Questionnaires based on IRT Model Tools

Applying IRT modelling for evaluating item and scale properties provide instrument developers a tool box of graphical techniques to improve their questionnaire's reliability, to shorten their questionnaires without sacrificing measurement precision, or to tailor their instrument for an individual or population.

For improving scale reliability, ICCs can be used to determine if one needs to modify the response category format, such as choosing between a five or seven-point Likert-type scale, to optimize the coverage of the item categories for measuring distinct portions of the construct continuum. Scale information functions, and equivalently the SEM curve, allow developers to determine where measurement gaps exist along the construct continuum suggesting the need to add new questions.

Developers who wish to reduce the length of their scale may use the item information curves or item-person maps to pick and choose the most informative set of items that cover the construct continuum they wish to measure. They can take that smaller subset of items and look at the new scale information curve to look at how information has changed based on the selected item set. Brief forms with comparable psychometric properties to its full versions are potentially useful for reducing respondent burden in busy clinics or long patient surveys.

Developers also can use information curves and item-person maps to tailor their instrument based on knowledge about the study population. For example, a developer may wish to tailor the Fear of Recurrence scale to be used as a screening or diagnostic tool to determine if a person may need psychiatric intervention or counselling. Information should be maximized for high fear levels along the θ continuum to provide accurate determination if a person falls beyond some predetermined cut off for medical attention.

To take full advantage of the powerful features that IRT modelling provides, applications of computerized-adaptive testing (CAT) allows a computer in real time to pick and choose the most informative subset of questions based on a person's response to previously administered questions. The methodological development of CATs is discussed by Drs. Chih-Hung Chang and Jakob Bjorner. The advantages of CAT-based assessment in health outcomes research and applications are featured in the papers written by Drs. David Cella and John Ware. Finally, the critical issues in developing and maintaining CATs for a variety of research and applied purposes are discussed in papers written by Drs. Colleen McHorney, Dennis Revicki, Jeff Sloan, Lawrence Fine, Robert O'Neal, and Laurie Burke.

Conclusions

This brief paper has focused on the graphical techniques made available by IRT modelling, which allow instrument developers to evaluate the properties of their scales and to revise the instrument resulting in a brief targeted questionnaire. There are an abundant amount of text-based information that is available in IRT software output files (such as fit statistics), and with experience interpreting these results, a researcher is armed with detailed information about their scale beyond simple item-total score correlations and mean scores.

IRT modeling is a complex methodology that in the right hands of a trained researcher can improve the questionnaires used in clinical and observational research. As will be discussed in Dr. Maria Orlando's paper, there are a number of assumptions that need to be made and one must consider such issues as which IRT model to use and importance of assessing model fit to the data. It must be emphasized that IRT is a statistical tool that cannot judge the content (or face) validity of a scale. It is important that a psychometrician work hand-in-hand with content experts throughout all phases of application from evaluating the assumption of unidimensionality to picking and choosing items. Also, IRT modeling should not replace descriptive methods used in classical test theory (CTT) but both CTT and IRT methods should be used to inform each other.

Acknowledgments

Thank you to Dr. Laurel Northouse, University of Michigan, and Dr. Suzanne Mellon, University of Detroit-Mercy, for sharing both the breast cancer data and the Fear of Recurrence questionnaire to use as illustrations for this paper. Also, thank you to Drs. Jakob Bjorner, Chih-Hung Chang, and Thomas Uttaro for reviewing a draft of this paper and providing excellent suggestions for improving the manuscript.

Key References to Learn More about the

Applications of IRT Modeling for Improving Patient-Reported Outcomes Measures

Embretson, S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates: Mahwah, NJ.

Reeve, B. B., & Mâsse, L. C. (2004). Item response theory modeling for questionnaire evaluation. S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.) *Methods for Testing and Evaluating Survey Questionnaires*. John Wiley & Sons, Inc.

Uttaro, T., & Lehman, A. (1999). Graded response modeling of the Quality of Life Interview. *Evaluation and Program Planning, 22*, 41-52.

Wilson, M. (2004). *Constructing Measures: An Item Response Modeling Approach*. Erlbaum Associates: Mahwah, NJ.

Figure 1: Item Characteristic Curves (a = 1.34, b_1 = -2.04, b_2 = -0.31, b_3 = 0.43, b_4 = 2.41). Parameter estimates derived from Samejima's Graded Response Model using MULTILOG software.

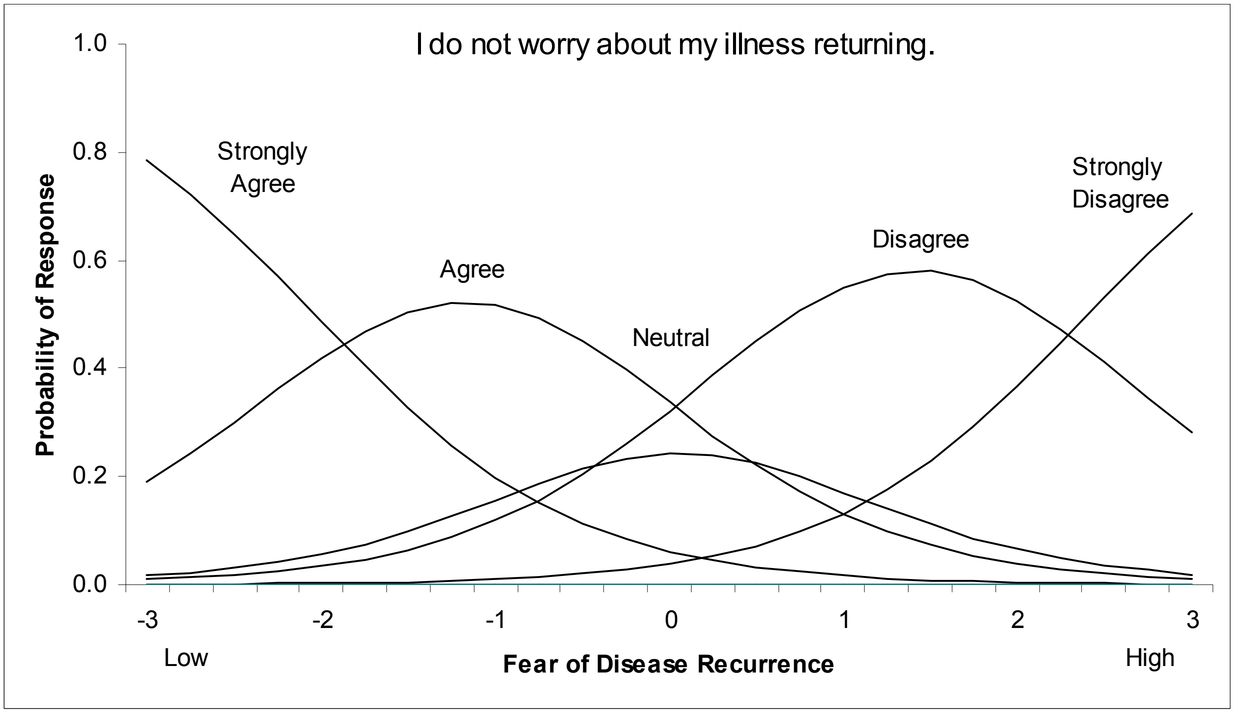


Figure 2: Item information curves for three items in the FOR scale. Curves derived from IRT parameter estimates from Samejima's Graded Response Model using MULTILOG software.

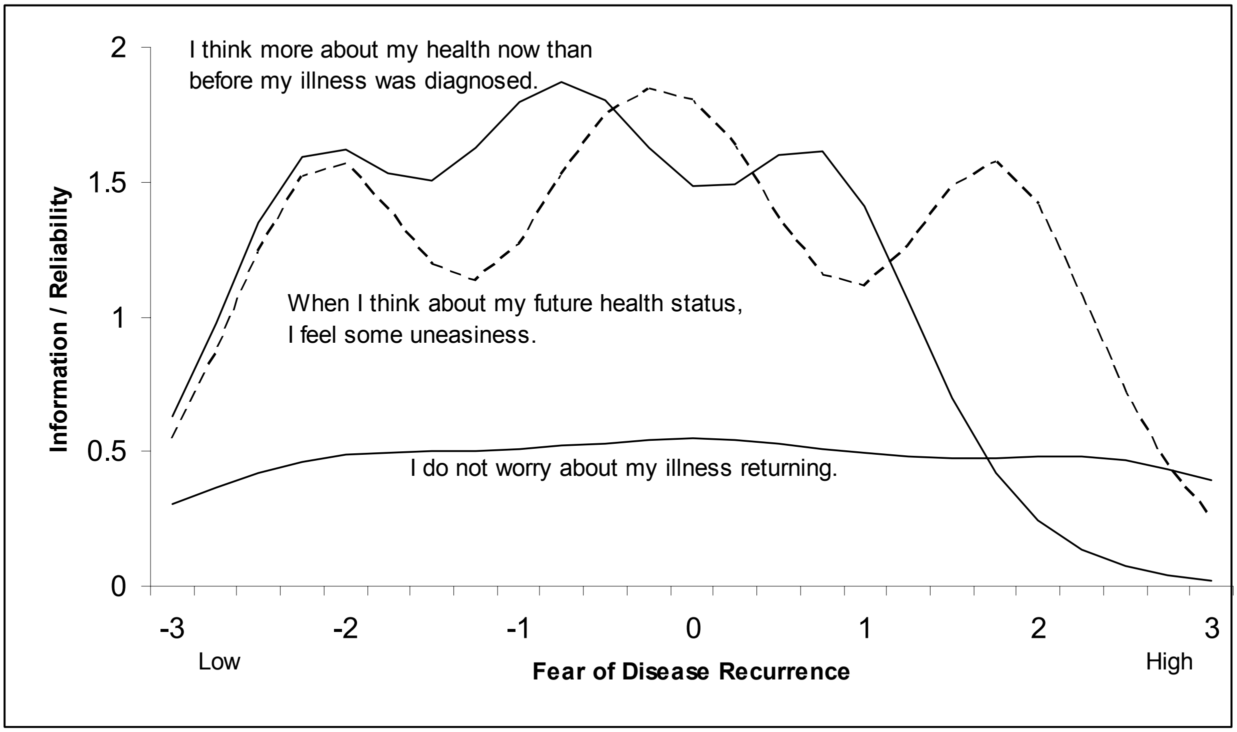


Figure 3: Item – Person Map (Wright Map) for the FOR Scale. Map based on parameter estimates from Masters Partial Credit Model implemented in WINSTEPS software

```
              <more>|<rare>
    4               +
                    |
                    |
                 X  |
                    |
                    |
                    |
    3               +
                    |
                    |
                XX  |
                    |
                    |
                    |
    2              T+
                    |
                    |
                 X  |
                    |
                XX  |  V20
               XXX  |
    1       XXXXXX S+
           XXXXXXX  |T
           XXXXXXX  |
       XXXXXXXXXXX  |  V8
            XXXXXX  |S V22    V5
          XXXXXXXX  |  V17    V19
           XXXXXXX  |  V21
    0 XXXXXXXXXXXX M+M V10    V16    V2     V4     V6
         XXXXXXXXX  |  V11    V12    V13    V15    V18    V9
           XXXXXXX  |  V7
      XXXXXXXXXXXX  |S
              XXXX  |  V1
               XXX  |  V3
               XXX S|T V14
   -1           XX  +
                 X  |
                    |
                XX  |
                    |
                 X  |
                 X T|
   -2            X  +
                XX  |
                    |
                    |
                    |
                    |
                 X  |
   -3               +
                    |
                 X  |
                    |
                    |
                    |
                    |
   -4               +
              <less>|<frequ>
```

Figure 4: FOR scale information curve. Curve estimated based on parameters derived by Samejima's Graded Response Model as implemented in MUTLILOG software.

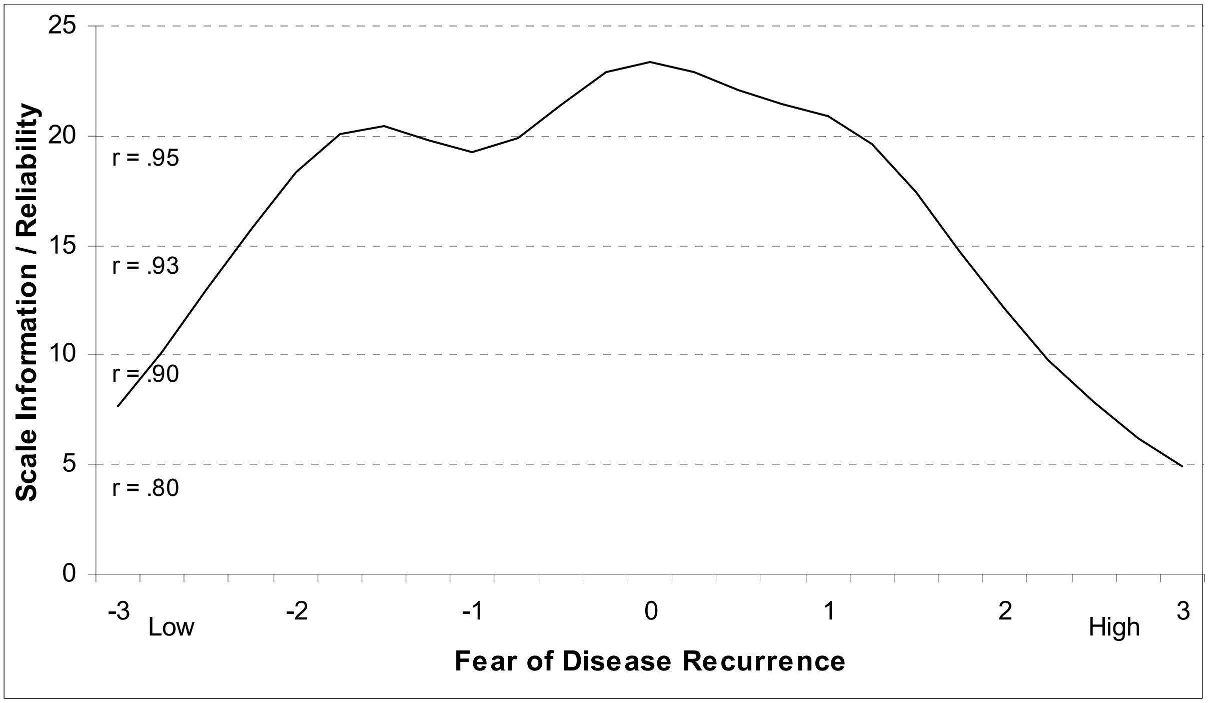
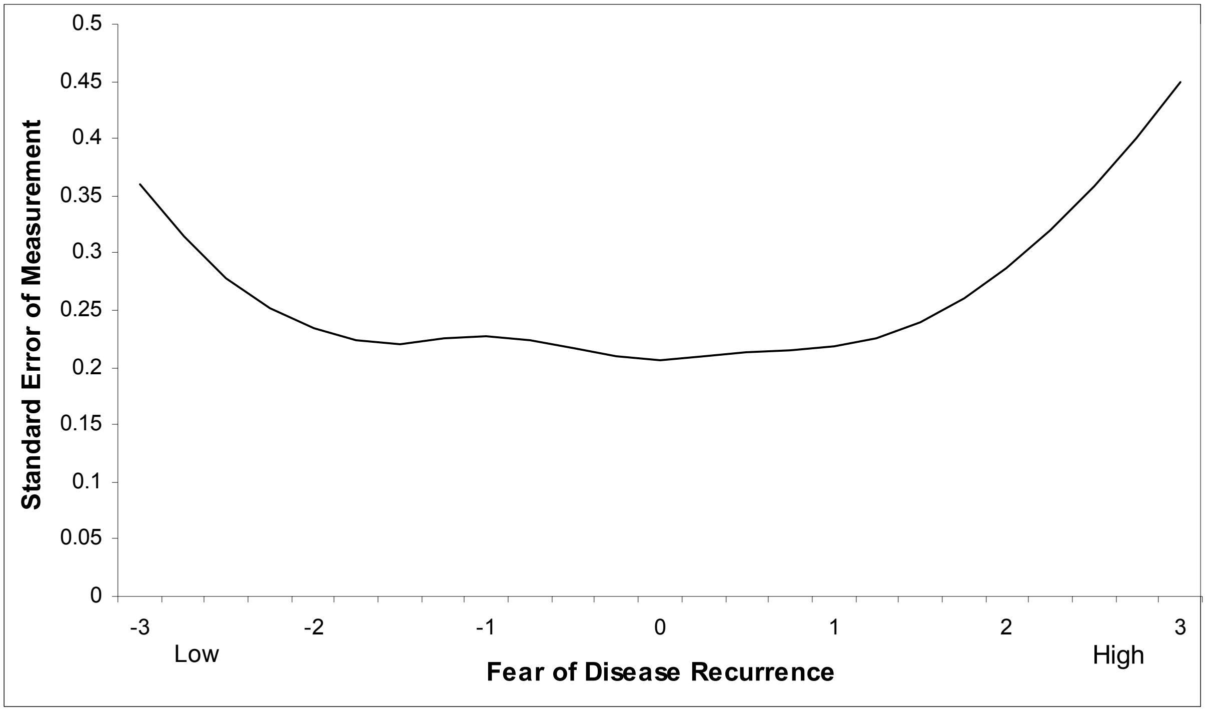


Figure 5: Standard Error of Measurment (SEM) Curve for FOR scale. Estimated Curve based on parameters derived by Samejima's Graded Response Model as implemented in MUTLILOG software.


Standard Error of Measurement
Low
High
Fear of Disease Recurrence