SURVEY DEVELOPMENT AND VALIDATION WITH THE RASCH MODEL

Kathy E. Green, University of Denver

and

Catherine G. Frantom, University of Missouri - Columbia

A paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, SC, November 14-17, 2002.

Abstract

Rasch analysis is illustrated using two examples: the first survey was not constructed using the Rasch model while the second survey was. Requirements of good measures are noted, Rasch statistical indices described and analysis illustrated in the context of two surveys, and benefits gained from Rasch analysis summarized. The paper closes by describing several remaining considerations in the area of measurement.

Introduction

The advancement of science depends on measurement. Developing good measures, however, can be challenging, particularly in areas where constructs are difficult to define. In survey research, for example, constructs may be ephemeral or ambiguous, and still must be assessed with maximum brevity. This and other measurement issues must be carefully considered as the development of theory is affected when measurement problems overwhelm the data. The presentation of uniform items to respondents of varied backgrounds without access to the redundancy present in spoken language yields a complex measurement problem. But, Wright (1977) said, "Progress marches on the invention of simple ways to handle complicated situations" (p. 97).

For the most part, measurement models found within item response theory (IRT) can provide the information needed to develop and/or assess the qualities of a desirable measure. A desirable measure is one that is simple and easy to use and is characterized by high quality of the information obtained--usually reported as reliability and validity. While some measures are relatively straightforward, like some of those used in the physical sciences, development of survey measures presents a number of challenges as follows: (1) Finding an optimal length for a survey can be difficult. In general, surveys should be shorter rather than longer - short instruments that maintain high quality are ideal. Although item redundancy adds to respondent burden and may increase unit and item nonresponse, decreasing instrument length adversely affects variance and thus impacts reliability and validity. (2) Identifying ways of effectively dealing with missing data requires, at minimum, a thorough understanding of processes involved in data collection. For instance, item level missing data plagues postal mail surveys, while forcing complete responses may lower the unit response rate to Web surveys. (3) Inter-item dependencies may present obstacles to constructing a measure. This has been evidenced in research on item order effects found for some item sequences (Converse & Presser, 1986) and complex reverse order effects in others (Sudman & Bradburn, 1982). (4) Identification of appropriate item and response scale functioning and changes in item and scale functioning across subpopulations or over time are critical to the accuracy of conclusions. (5) The data collected through administration of the instrument should be capable of meeting criteria for statistical analyses.

A good measure should yield invariant scores. Invariance describes the ‘scope of use’ properties of a measure. For example, a ruler provides a measure of height in inches. The ‘height’ scores are invariant: regardless of the ruler used, a person’s height remains constant and the ruler can be used with anyone. A ruler’s use is not restricted to particular groups of people and is not biased towards men or women. Arguably, different item response patterns can provide interesting information about the characteristics or culture of respondents. For example, patients with some forms of osteoarthritis may have more difficulty agreeing with items that reflect pain in gripping objects than patients with rheumatoid arthritis. This gives us information about the different patient groups. However, the failure of invariance prohibits group comparisons since the variable’s definition changes for the different types of patients. This is a validity issue. Though there are likely few measures outside the physical sciences where invariance is demonstrated, it is a worthy goal.

Specific objectivity is another desirable characteristic in a measure. Specific objectivity means that a person’s trait is independent of the specific set of items used to measure it. For example, it shouldn't matter *which* ruler is used to measure a person’s height; any ruler could be used and any one used would be independent of the person’s height. Additionally, a measure with specific objectivity would not be affected by missing data. Hence, despite missing data, the measure would still be useful and provide credible information. Measures with specific objectivity can be tailored to any given respondent, thus permitting individually administered surveys and precluding administration of items that are not appropriate for a particular respondent.

A statistic known as ‘fit’ provides an internal mechanism for identifying inappropriate responses to the items, allowing exclusion or re-assessment of persons whose responses make no sense, i.e., do not fit, according to our understanding of the construct. For instance, our understanding of depression as a construct should be reflected in the pattern of participants' responses. A person who is more depressed would be expected to agree more strongly with items on a depression survey than someone who is less depressed. Questions should be raised when logic of the construct does not prevail, as when a person agrees with an item on suicidality but not with an item on feeling sad. In this case, the person’s understanding of the construct, his sincerity, and our own understanding of the construct should be examined. Thus, ‘fit’ provides an index of the degree to which responses conform to a logical pattern as well as an indication of the measure’s validity for a specific individual. Similarly, ‘fit’ permits assessment of the validity of the overall measure by providing a means to identify poorly functioning and/or biased items. Item fit is an index of how well items function in reflection of the trait. Items with an acceptable fit index, i.e., that fit better, are more useful in measuring a trait than items that fit poorly.

In addition to examining fit, a participants' use of a response scale (e.g., strongly disagree to strongly agree, yes-no-maybe) can be understood using item response theory. The consensus in scale design regarding the ideal number of response categories to use seems to be 4-7 categories with strongest support for 5 choices. Logic dictates that an individual responding with a higher category selection (strongly agree, for example) would have more of the characteristic being measured than someone responding with a lower category selection. Once again, if logic does not prevail, understanding of the scale and the validity of the person's responses should be questioned.

As already stated, IRT models are useful for assessing the properties of an instrument. They can, in fact, effectively provide the benefits listed above. They fail, however, in terms of ease of use. Knowledge of specific software and familiarity with output interpretation must be acquired. Though this prospect may be daunting, the outcome makes it worth the effort.

The next section provides a conceptual overview of the information that can be obtained from the Rasch model (Rasch, 1960). This is perhaps the simplest and most useful of the IRT models. Examples of application of the Rasch model to specific instruments follow. The instruments are: (1) the Separation-Individual Test for Adolescents (Levine, Green, & Millon, 1986), and (2) the Physician Perceptions Survey (Frantom, 2001). The former instrument was not created with the Rasch model in mind while the latter was. Finally, the benefits gained from use of the Rasch model are summarized along with some of the measurement challenges remaining in survey research.

Overview of Rasch Model Indices

The Rasch model is a mathematical formula that specifies the form of the relationship between persons and the items that operationalize one trait. Specifically, the likelihood of higher scores increases as people have more of the trait and decreases as they have less of the trait, whereby items become more difficult to endorse. The Rasch model assumes that item responses are governed by a person's position on the underlying trait and item difficulty. As implied by the theory's name, *item* responses are modeled rather than sum total responses. The model makes no allowance for deliberate or unconscious deception, guessing, or any other variable that might impinge on the responses provided. We model only the trait and not minor, peripheral influences. The Rasch model is a one-parameter model, meaning that it models the "one" parameter difference between person position and item difficulty. However, it actually provides two parameter estimates: person position and item difficulty, also referred to as person logit and item logit respectively, where a logit is a translation of the raw score. Equal-interval measures can be constructed using the Rasch model, where persons and items exist on a common scale. In other words, raw scores are nonlinearly transformed into position estimates for items and persons so that the data best fit the model.

Fit statistics provide the indices of fit of the data to the model and usefulness of the measure. Fit statistics include the average fit (mean square and standardized) of persons and items, and fit statistics reflecting the appropriateness of rating scale category use. The fit statistics are calculated by differencing each pair of observed and model-expected responses, squaring the differences, summing over all pairs, averaging, and standardizing to approximate a unit normal (z) distribution. The expected values of the mean square and standardized fit indices are 1.0 and 0.0, respectively, if the data fit the model. Fit is expressed as "infit" (weighted by the distance between the person position and item difficulty) and as "outfit" (an unweighted measure). Infit is less sensitive than outfit to extreme responses.

Person fit to the Rasch model is an index of whether individuals are responding to items in a consistent manner or if responses are idiosyncratic or erratic. Responses may fail to be consistent when people are bored and inattentive to the task, when they are confused, or when an item evokes an unusually salient response from an individual. Similarly, item fit is an index of whether items function logically and provide a continuum useful for all respondents. An item may "misfit" because it is too complex, confusing, or because it actually measures a different construct (e.g., locus of control rather than depression). As survey researchers, we work to use the same language that respondents use and carefully frame items in that language on the survey. Fit statistics allow us to check whether we truly have a basis for communication. As already noted, “fit” is expressed as a mean square and as a standardized value. Values for differentiating "fit" and "misfit" are arbitrary and should be sufficiently flexible to allow for researcher judgment. Also, some fit values will appear too large or too small by chance.

Person and item separation and reliability of separation assess instrument spread across the trait continuum. “Separation” measures the spread of both items and persons in standard error units. It can be thought of as the number of levels into which the sample of items and persons can be separated. For an instrument to be useful, separation should exceed 1.0, with higher values of separation representing greater spread of items and persons along a continuum. Lower values of separation indicate redundancy in the items and less variability of persons on the trait. To operationalize a variable with items, each item should mark a different amount of the trait, as for instance, the way marks on a ruler form a measure of length. Separation, in turn, determines reliability. Higher separation in concert with variance in person or item position yields higher reliability. Reliability of person separation is conceptually equivalent to Cronbach's alpha, though the formulas are different.

Rating categories within items should form a continuum of less to more. That is, endorsing a lower category should represent being lower on the trait, e.g., “a little like me,” than endorsing a higher category, which would be, e.g., “a lot like me”. Lack of order in rating scale categories suggests a lack of common understanding of use of the rating scale between the researcher and the participants. Inconsistent use of the rating scale affects item fit and placement. Such problems can be resolved, albeit post hoc, by combining categories and reanalyzing the data to reassess the optimal number of categories for that data.

Persons and items can "overfit" or "underfit." Overfit is indicated by a mean square value lower than 1.0, and a *negative* standardized fit. Overfit is interpreted as too little variation in the response pattern, perhaps indicating the presence of redundant items. Although it provides a guide to refining an instrument, it is otherwise probably of little concern. Underfit (“noise”) is indicated by a mean square >1.2 and standardized fit >2.0 (as one rule of thumb) and suggests unusual and/or inappropriate response patterns. These indices can be used to identify and sometimes correct a measurement disturbance.

Placement of items and persons on a common scale permits evaluation of scale function relative to the sample. The Rasch model software, WINSTEPS (Linacre & Wright, 1999-2001) can graph person position with item position. Simultaneous positioning of items and person responses illustrates where responses place each person with respect to those items. This graph is useful in three ways: (1) It can be used to determine the extent to which item positions match person positions. If the positions do not line up, the items are likely inappropriate for the persons (e.g., too easy or too hard to agree with). (2) Gaps in the measure can be detected, suggesting where items might be added. (3) Item order can be reviewed to assess the validity of the measure. When considering the construct of depression, we might anticipate that items about physically-oriented complaints such as difficulty with sleeping or eating will be easier to endorse, and easier to endorse more strongly, than psychological problems such as hopelessness or guilt. Logic in the arrangement of items indicates that a researcher understood the construct, adequately operationalized it with the items written, and successfully communicated it to respondents via the items written to define it. The logic of item placement depends on qualitative judgment by the researcher, and is based on his or her knowledge of and experience with the construct.

For more information regarding Rasch model statistics and their utility in research, see Fox and Jones (1998), Bode and Wright (1999), or Snyder and Sheehan (1992). For a lengthier explanation of the Rasch model, see Bond and Fox (2001) or Wright and Stone (1979).

Data Requirements for Design and Analysis with the Rasch Model

An instrument can be developed using classical test theory and/or item response theory. In general, the tasks involved are the same. Using the Rasch model, however, provides an opportunity to attend to the anticipated item positions along a continuum of item endorsement difficulty. A panel of experts can be a valuable resource for judging the difficulty level of items through a sorting process. A hierarchical ordering of items by the panel of experts that is similar to the ordering determined by the primary researchers would suggest that they have a common understanding of the construct. The empirical item order would be expected to conform to a similar pattern. An instrument best defines a trait when the items written to support it function consistently throughout the instrument development process. Inconsistencies can suggest areas for reconsideration. Note that data collected from instruments that were *not* designed with Rasch analysis in mind can still utilize the Rasch model trait continuum to see how well the construct was understood. An initial requirement, then, is item sorting by the primary researcher and an expert panel.

A sample size of at least 100 and a minimum of at least 20 items are suggested for obtaining stable indices when using Rasch analysis. Analyses can still be conducted, however, with far fewer people and items. (See, for example, Smith, Lawless, Curda and Curda, 1999, detailing results of an analysis with 48 people and 14 items.) Arguably, Rasch analysis can be informative with small samples since the data are responses to individual items which generates an N x n matrix (Persons x Items).

The Rasch model can be used with categorical data, rating scale data, or frequency count data. Logit person and item positions can be adjusted for the influence of extraneous variables (e.g., judge severity, person gender).

The following section illustrates the use of Rasch model indices in the context of (1) an instrument in adolescent development, the Separation-Individuation Test of Adolescence (SITA: Levine, Green, & Millon, 1986) that was *not* developed with Rasch analysis in mind, and (2) the Physician Perceptions Survey (Frantom, 2001), an instrument that *was* developed with Rasch analysis in mind.

Illustrative Rasch Model Analyses

The SITA: Practicing Mirroring

The SITA was developed by Levine, Green, and Millon (1986) to measure six dimensions of adolescent separation-individuation derived from Mahler's (1968) psychodynamic and object relations theory. This theory-based instrument was factor analyzed by the original authors and its structure re-examined by Levine and Saintonge (1993) and Kroger and Green (1994). The 103-item, self-report instrument was revised by Levine and Saintonge to assess nine dimensions. All authors found moderate support for the structure, reliability, and validity of the instrument, with some subscales evidencing stronger convergence, higher reliability, and higher correlations with external instruments than others. The SITA was intended for use as both a research and clinical instrument relevant to adolescent interpersonal relationships. The SITA was selected as an exemplar instrument because of its use in research, because it is theory-based, and because its structure has been examined in the past and has been supported. The 15-item subscale, Practicing Mirroring, achieved the highest internal consistency and highest correlation with a validation measure in Levine and Saintonge's study. This subscale was used here to illustrate Rasch analysis.

Method

Research participants were 131 late adolescent New Zealand university students (84 female, 47 male) with mean ages of 19.8 (females) and 19.2 years (males). All SITA questionnaires were administered individually as part of a study of identify formation during late adolescence. Further detail about the sample and data collection procedures can be found in Kroger and Green (1994). The Practicing Mirroring scale assesses the degree of narcissism experienced by the respondent. It has 15 items and internal consistency estimates of .85-.87., and has emerged clearly in past factor analyses.

Results

Prior to interpretation of the item and person logit (position) scores from a Rasch analysis, appraisal of whether the data fit the model reasonably well is required. Table 1 presents overall information about whether the data showed acceptable fit to the model. The mean infit and outfit for person and item mean squares are expected to be 1.0. For these data, they are all .99. The mean standardized infit and outfit are expected to be 0.0. Here they are -.3 for persons and -.4 for items. The items overfit, on average. This suggests that the data fit the model somewhat better than we would expect which may signal some redundancy--possibly redundant items. Redundancy gives us an indication of how we may trim items to reduce the length of the instrument. The standard deviation of the standardized infit is an index of overall misfit for persons and items (Bode & Wright, 1999). Using 2.0 as a cut-off criterion, both persons (standardized infit standard deviation = .56) and items (standardized infit standard deviation = .38) show little overall misfit. Here the data evidence acceptable fit overall.

The next overall statistic we look at is called separation, the index of spread of the person positions or item positions. For persons, separation is 2.30 for the data at hand (*real*), and is 2.56 when the data have no misfit to the model (*model*). This suggests we have measured persons on a continuum rather than a dichotomy. If separation is 1.0 or below, the items may not have sufficient breadth in position. In that case, we may wish to reconsider what having less and more of the trait means in terms of items agreed or disagreed with, and on revision, add items that cover a broader range. An exception to this occurs if we are using a test to make dichotomous decisions. That is, if we need to decide "pass" or "fail," "likely brain injury" or "unlikely brain injury." In that situation, we want items that center on what we define as a cut-off point. But, in such a case, we are creating a categorization system, not a measure. Item separation for the present case is 4.51, an even broader continuum than for persons. It is typical to find larger separation values for items than for persons, a function of the fact that we often work with a small number of items and a larger number of people, here 15 items and 131 people. Separation is affected by sample size, as are fit indices and error estimates. With larger sample sizes, separation tends to increase and error decrease.

The person separation reliability estimate for these data is .84. We are given a conceptual analog to person reliability in item reliability--which estimates internal consistency of persons rather than items (here, .95).

Note that the mean for items is 0.0. The mean of the item logit position is always arbitrarily set at 0.0, similar to a standard (*z*) score. The person mean here is -.36, which suggests these items were difficult, on average, for persons to agree with but were fairly well matched to the perceptions of the sample. If the person mean is positive, the items would, on average, be easy to agree with. The persons would have a higher level of the trait than the items do. If the person mean were -1, -2, or +1 or +2, we would consider the items potentially too hard or too easy for the sample and might seek a different test.

Table 2 and Figure 1 display how the response scale was used. For these data, the response scale was 1 (never true for me) to 5 (always true for me). Table 2 lists the step logit position, where a step marks the transition from one rating scale category to the next, e.g., from a 4 to a 5. "Observed Count" is the number of times the category was selected across all items and all persons. "Observed Average" is the average of logit positions modeled in the category. It should (and does) increase by category value. Persons responding with a "1" have an average measure (-1.72) lower than those responding with a "2" (average measure = -1.05), etc. There is no substantial misfit for categories 1 through 4, though the misfit for category 5 is a little higher than one might wish (mean square misfit >1.2), suggesting too much noise in responses of "5." "Sample expected" is the optimum value of the average logit position for these data. Sample expected values should not be highly discrepant from observed average--and for these data they are not. Infit and outfit mean squares are each expected to equal 1.0. Step calibration is the logit calibrated difficulty of the step. These values are expected to increase with category value, which they do. The step standard error is a measure of uncertainty around the step calibration. Another view of step function is the Thurstone threshold, which is the location of the medians--where the point of observing the categories below equals the probability of observing the categories equal to or above that point. A final way of examining step use is via probability curves (Figure 1). These curves display the likelihood of category selection (y-axis) by the person-minus-item measure (x-axis). If the difference in logit position between the person and item is +1.0, while any response is possible, the most likely category response is 3, followed closely by a 4. If the difference in logit position between the person and item is more than 2.56 the most likely response is a 5. If all categories are utilized, each category value will be the most likely at some point on the continuum, as is the case here, and there will be no category inversions where a higher category is *more* likely at a lower point than a lower category. (This is not the case here, but would be, for example, if the 3's and the 4's switched places.) The transition points between one category and the next are the step calibration values from Table 2. For these data, all categories are being used and are behaving according to expectation.

At this point in our analysis we know rating scale categories were used appropriately and the data overall fit the model. We then proceed to examine persons and item fit to see how individual items functioned and how individual persons responded to the items. If we see evidence of inappropriate use of the scale at this point, we would need to try collapsing categories to see whether that would remedy the problem. For example, if there were an inversion of 4 and 5, we could collapse those two categories into one and examine the structure of the collapsed scale by rerunning the analysis. Linacre (2002) provided the following criteria for deciding when categories might be collapsed with adjacent categories. (1) Categories have fewer than 10 responses (underused), (2) categories are infrequently used in comparison to other categories, (3) average and estimated step calibrations are dissimilar, (4) categories are disordered, (5) steps are not sufficiently different (e.g., <1.4 logits apart) or are too different (e.g., >5 logits apart), and (6) outfit mean square exceeds 2.0.

Table 3 displays the items in order of worst to best fitting. Entry number is the item's location in this scale of 103 items. Raw score is the total number of "points" the item got across the entire sample. Count tells us that of 131 participants, all responded to all items, not surprising since the instrument was individually administered. Measure is the logit position of the item, with error being the standard error of measurement for the item. Here error is the same for all items (.12). Generally this is not the case. No definitive rules exist regarding what is considered acceptable and unacceptable fit but some suggestions for acceptable fit are as follows: (1) Mean square (infit or outfit) between .6 and 1.4 , (2) mean square (infit or outfit) between .8 and 1.2 (Bode & Wright, 1999), (3) mean square less than 1.3 for samples less than 500, 1.2 for 500-1,000, and 1.1 if n>1,000 (Smith, Schumacker, & Bush, 1995), (4) standardized fit (infit or outfit) between -2 and +2, (5) standardized fit between -3 and +2, and (6) standardized fit less than +2 (Smith, 1992). Infit is a weighted goodness-of-fit statistic, where unexpected responses to items close to the person's logit position are weighted more heavily than unexpected responses to items far away from the person's level (*in*formation laden). Outfit is unweighted and so is sensitive to extreme unexpected responses (*out*lier sensitive). The score correlation is the correlation between item score and the measure (as distinct from a total score), and so is an item discrimination index. It should be positive. For these data, it appears that the item in the 87th position, "physical appearance attractive" does not fit well with the rest of the scale as its mean square and standardized infit *and* outfit values exceed all recommendations. In scale revision, we might consider rephrasing or deleting this item. The rest of the items seem to fit, perhaps too well. According to some fit criteria, we might examine items in the 43rd, 88th, and 97th positions, as these have less random fluctuation in responses than expected. These items are called overfitting items. If an item fits substantially better than the items as a set, it may be too sharply discriminating. Items that have a very high score correlation may actually not fit with the rest of the set of items. A mark on a ruler that differentiates 3.10 centimeters from 3.11 centimeters (high discrimination) is not useful in tailoring a suit, whereas a mark for 3.1 versus 3.2 is.

Table 4 displays the responses of one person, Person 4, to these 15 items, with an integer standardized residual showing which responses were unexpected and how unexpected they were. This person's logit position was -.54, meaning it tended to be difficult for this person to agree with the items--his score suggested him to be a little less narcissistic than the average of -.36. This person unexpectedly said items 32, 83, and 92 were always true for him (category 5), resulting in positive residuals of 2, 3, and 3. We expected a higher category selection for items 87 and 97, with residuals of -3 and -2, respectively. No single residual is extremely high but rather misfit is spread across 5 of 15 items. This person shows an unusual pattern of response and we may wish to re-interview him or refer him for further examination. His overall infit and outfit at 3.9 were high. If the instrument were used clinically, we would want to get additional information about this person. If we were *developing* an instrument, however, we would probably simply discard his data as adding noise to our work.

At this point, if we have identified problems with scale use, particular items, or particular persons, we would address these problems by collapsing scale categories, deleting items or persons' data, and then rerun the analysis and begin our interpretation again.

Possibly the heart of the Rasch analysis is provided in Figure 2, the map of persons and items displayed in tandem. The distribution of person positions is on the left side of the vertical line and items on the right. Each "X" represents two persons in this figure. "M" marks the person and item mean, "S" is one standard deviation away from the mean, and "T" is two standard deviations away from the mean. Those at the upper end of the scale agreed with more items and agreed more strongly. The distribution of person logit positions is positively skewed. In Figure 2, we can see that there are numerous persons whose position is *below* where items are measuring--there are no items that match these persons' levels of the trait very well. We see that the items cover a range of -1 to +1 logits in difficulty, narrower than the range of about -2.5 to +4.0 for persons. As such, we may try on scale revision to write easier and harder items to extend the range of the trait measured. Also, note that at one point on the scale there are 4 items at the same position. We may consider dropping one or two of them as redundant. Finally, the key question is--does the order of items make sense? Should people find it harder to agree that they "feel powerful" (item 2) than that they are "better off than most" (item 32)? Would someone who is more narcissistic agree that "others admire me" (item 88) while someone less narcissistic disagree with this but agree he "acts like a leader" (item 18)? Note also the gap between items 18 and 91. On instrument revision, items should be created that mark that level of the trait. The ability to structure items that likely mark intermediate levels of the trait requires thorough understanding of the trait and considerable ability to communicate through written items. Filling gaps in the item distribution is by no means an elementary task.

Additional information generated by WINSTEPS (Linacre & Wright, 1999-2001) but not presented here includes a table of the raw score-to-logit position conversion (e.g., a raw score of 53 corresponds to a logit position of .98). This is the table that provides the ordinal to interval score conversion. Convergence information is provided as well. If the program takes many (e.g., 100) iterations to converge--to arrive at acceptable, stable parameter estimates--the data may possess some characteristics that make it unsuitable for analysis. These data converged in 15 iterations. The convergence table lists which persons, items, and categories caused the most difficulty, e.g., Iteration 3--item 13, person 107, category 3; Iteration 12--item 13, person 41, category 3.

An output file with calibrated person positions (in logits) can be generated. These values are on an interval scale and can be used in subsequent statistical analyses. Each item and person position estimate is accompanied by an error estimate. Estimates are possible for all persons (except those with extreme responses) regardless of the amount of missing data, as long as we have *some* usable responses. But, if there are little data upon which the estimate is based, the error estimate will be larger. Thus missing data are accommodated but not ignored.

Physician Perception Survey

The Physician Perception Survey (PPS) was developed to test a theory of physicians' perceptions of patient autonomy. Evidence and myths pertaining to physician behaviors within the medical culture (e.g., Anderson & Zimmerman, 1993; Blumenthal, 1994; Kauffman, 1983) generated the hypothesis that physician paternalism and perfectionism would impact perceptions of patients' abilities to make choices in health care. Thus, a primary goal of the development study was to create an instrument that could address the following questions: Do physician paternalism and perfectionism exist, and if so, how might they influence perceptions of patient autonomy in terms of patients' rights to make independent and informed decisions?

In brief, paternalism was defined as the act of treating others in a fatherly way via a position of superiority. Though the concept of paternalism is found throughout the medical literature, no true measures of the construct could be found. Perfectionism refers to a doctor's belief that he/she possesses superior epistemic knowledge that ultimately leads to a stance of "doctor knows best." Perfectionism is most often defined in terms of setting excessively high standards for one's behaviors (e.g., Frost, Marten, Lahart, & Rosenblate, 1990; Pacht, 1984), tendencies for overly critical evaluations of one's own behavior, and over-concern for mistakes (Frost et al.). Though the project was initiated from the literature on paternalism and perfectionism, these aspects of physicians' perceptions were labeled "Physician Disclosure" and "Patient Needs." The PPS was selected as an exemplar instrument because it was designed with Rasch analysis in mind.

Method

Instrument development followed the guidelines of Benson and Clark (1982) and DeVellis (1991), beginning with broad conceptualizations of the constructs and culminating with specific descriptions of factor-derived subscales. Developing an instrument is an iterative process requiring review of data on test and item levels. Though the process cannot be truly considered linear, decisions are made at various steps along the way based on predetermined statistical criteria and judgments about qualitative aspects of the items themselves. Classical and item response theories were used for instrument development and analysis. Development combined principal component analysis based in classical test theory to decide the structure for the final revised instrument, and simultaneous review of item function through Rasch analysis to assess construct coverage and order. Decisions to retain or delete individual items were made based on inter-item and item-scale relationships as tracked by Rasch model statistics, e.g., fit, separation, distribution of items, step measures; and principal component analysis using factor loadings and eigenvalues. Reliability checks were integrated with the primary analyses to aid in decision-making.

A pool of 67 initial items was generated based on the proposed theory. Instrument development began with an expert review of the items. A 4-point response scale was used, with ratings ranging from 1 (strongly disagree) to 4 (strongly agree), with no neutral response option provided. Scoring was designed so that higher subscale scores would reflect more of the trait being measured (i.e., paternalism or perfectionism), and as such, a higher total score would theoretically describe a 'lesser perception' of patient autonomy. The instrument evolved through a pilot administration of 42 items, reduction and a second administration of 21 items, and identification of the final 15-item instrument comprising two somewhat distinct subscales. (A 12-item scale was ultimately proposed as an alternative single scale solution. See Frantom, 2001, for detailed information on the results at each step of measure development.) The expert review and the field administration results are described below.

A panel comprising two experienced medical ethicists, two practicing physicians, and two psychometricians reviewed the pool of 67 items with instructions to assess their face and content validity, evaluate the relevance of the items to the subscale they proposed to measure, order items (with a card sort) in terms of difficulty level (easy, medium, difficult), and judge items for clarity and conciseness. The goal was to obtain 40-50 items for the pilot instrument.

In the field administration, 500 were surveys distributed to U. S. physicians across specialty areas; 167 were returned yielding a 33% response rate from the Midwest (87%) and regions outside the Midwest (13%). The sample consisted of physicians who were approximately equally male (55%) and female (45%), predominantly Caucasian (79%), and practicing in a teaching hospital (61%). The overall mean age of physicians was 41.43 (SD=12.45) who averaged 14.26 (SD=12.39) years out of medical school, and on average had been practicing at their current setting for 7.16 (SD=8.40) years.

Subscales were identified using exploratory factor analysis and were individually analyzed using Rasch analysis. At each step, data, targeting, and response structure were checked for fit to the Rasch model. Adjusted standard deviations and person separation and reliability statistics were tracked through an iterative process, where items were added and/or deleted respective to their influence on the individual subscales as determined by mean square values for infit and outfit. Reliability of person separation was used to demonstrate whether respondents were being adequately separated by items along the continuum representing the construct, as well as provide an indication of replicability for person placement across other items measuring the same construct. Similar to Cronbach's alpha, perfect reliability would be 1.0 and random data would generate a relationship of 0.0. According to Wright (1977), a variable is "sharpened" as the adjusted standard deviation increases. Thus, a "good" instrument would be one with a high, adjusted standard deviation and a person separation of at least 1.0. An increase in adjusted standard deviation and person separation values was sought for both subscales with each iteration of the analysis.

Results

Expert Review. Item quality and content relevance were determined based on reviewers' matching of individual items with either perfectionism or paternalism and their judgment of whether the item related overall to patient autonomy. Decisions to retain items for the pilot study were made based on the following criteria: (1) Items received at least 3 votes for either paternalism or perfectionism *and* (2) at least 4 votes for being related to patient autonomy *and* (3) few or no votes in the 'neither' category. Table 5 contains an example of how results were tallied, listing the first 10 of the initial 67 items. "Patient Autonomy" in Table 5 refers to whether reviewers judged the content of the item as being related to patient autonomy in general. Overall, the expert review resulted in retention of 41 out of 67 original items for the pilot study.

Item difficulty ratings were tallied for the retained items to determine coverage of the construct. Items were classified as being easy, moderate, or difficult to agree with if they received at least three votes for one particular difficulty level. Seven items were judged as easy, 24 as moderate, and 14 as difficult. (Four items received 3 votes each for two difficulty categories.) In general, reviewers agreed that the distribution of 41 items adequately covered the construct.

Field Administration. First, suitability of the subscales for Rasch analysis was evaluated by reviewing overall fit and separation indices separately for each factor-based subscale (Tables 6 and 7). For both persons and items for both subscales, overall fit indices indicated sufficiently good overall fit to the Rasch model to allow item and person evaluation. Note that person separation for both subscales just exceeded 1.0, indicating that items cover a rather narrow range of the trait continuum.

The average measure of steps for both subscales indicated movement in the expected direction and step measures suggested good use of the 4-point scale. Figure 3 provides step probability curves for 'Physician Disclosure.'

Item and person fit indices were then reviewed (Tables 8 and 9), with reasonable fit for items. No person data were deleted due to severe misfit.

A "good" test shows an even spread of items along the variable void of gaps and targeted to person ability (Wright, 1977). Figure 4 shows the distribution of the persons and items for ‘physician disclosure’. The ordering of items is such that a person who agrees with a more difficult item should also agree with an easier one. The item-person map for the 'physician disclosure' subscale illustrates that the average item position was greater than the average person position resulting in some misalignment of persons and items. Gaps in the coverage of the construct are found on the lower end of the continuum where there are people and no items. Item content was reviewed to determine if item hierarchy spelled out a meaningful construct from easiest to agree with at the bottom to hardest at the top. It does appear that the more ethically complex items are positioned on the upper end of the continuum as expected and as initially rated by the expert reviewers.

The same process was used to analyze the items that defined the second subscale, 'patient needs.' Interestingly, after a number of iterations, removing and re-entering items, items on the shorter ‘patient needs’ subscale functioned better than those on the ‘physician disclosure’ subscale. For ‘patient needs,' person separation was low, but fit was good with a reasonable spread of items and perfect alignment of items with persons, suggesting an appropriate level of difficulty for the sample (Figure 5). The perfect matching between early expert reviewer difficulty ratings and the item level difficulty determined by the data is noteworthy. This finding lends credibility to the expert reviewers' judgment of the level of ethical complexity of the items, and hence, to the raters' reflected overall knowledge of the construct.

The person separation and reliability of this subscale could be improved with additional items to increase spread. 'Physician disclosure' also had low reliability of person separation, and fit was only fair, with gaps in coverage on the easy end of the continuum resulting in misalignment of items and persons. Though more items would be needed to improve reliability and coverage, in general, the items retained for each subscale functioned reasonably well.

Limitations of the Rasch Model in Practice

The Rasch model is termed a "strong" model since its assumptions are more difficult to meet than those of classical test theory. The benefits derived from its use come at the cost of, on occasion, failure to define a measure at all. When data do not adequately fit the model, the instrument construction process must begin anew. An overall failure could occur if items are poorly constructed or are not comprehensible to the population, if there is a blatant mismatch between the respondent group's abilities/attitudes and item difficulties, or there are anomalies in the item-person interactions. In some cases, the data may adequately fit the model overall without defining a continuum on a trait. Items that vary little in level of difficulty may fail to define a trait, even though they may result in high internal consistency reliability estimates.

Invariance may fail across subgroups in a sample. While this does not invalidate the measure, it limits it use. Invariance may fail when items have different meanings depending on respondents' gender, ethnicity, socioeconomic status, or other variables. It may also fail due to order effects, item phrasing (such as negation), survey formatting, or other contextual variables.

Finally, a major limitation of use of the Rasch model in survey research is the widespread use of single-item indicators. To assess a trait, multiple items are required. Although demographic, concrete information can be gathered from single-item responses, constructs tend to be more complex in nature.

Conclusions and Measurement Challenges

The Rasch model offers much more information to the test developer and user than classical test theory. It offers the survey researcher the opportunity to clarify measured constructs, to "bank" items with calibrated logit positions, to assess item functioning across groups, to determine whether items are flawed, and to find whether respondents are providing answers that correspond to the researcher's view of rational behavior. Beck and Gable (2001) state that the adequacy with which the attitude continuum is assessed is a critical function of IRT. The hierarchy of difficulty displayed is critical to creation of a scale with well-spaced items that covers a substantial length of the construct. The Rasch model can also deal well with the missing data that plague survey research. There are, however, additional areas for exploration.

Fit indices in the Rasch model have been criticized as being influenced by the testing context, making definition of a standard for defining fit versus misfit difficult. Varied standards exist for deciding fit to be adequate or not, and for deciding on deletion of item or person data. A fit standard is also absent for item analyses conducted using classical test theory. Karabatsos (2000) suggests that more investigation of the characteristics of fit indices be conducted.

People who respond at the extremes on a measure (e.g., strongly agree with all items, get all items wrong) are not scalable with the Rasch model and their responses are not used in calibration of items. This may not be a problem diagnostically but it is inconvenient in terms of score reporting. Perhaps a convention for reporting results falling at the extremes of a measure could be established. The Rasch software, WINSTEPS (Linacre & Wright, 1999-2001), assigns a logit position of 1.00 logit above (or below) the highest (lowest) person position to those persons with extreme positive or negative measures.

The measurement of change has long been problematic. In measuring change the stability of the construct should be investigated as reflected by a set of items and changes in use of the scale over time, as well as shifts in individuals' positions on the construct. The method for doing this is convoluted and does not readily apply to multidimensional instruments. Further, when differences in conclusions appear across methods of analysis, it is not clear which is correct since controlled simulation studies are lacking (Wolfe & Chiu, 1999). The Rasch model assumes a unidimensional trait. With complex instruments that assess multiple constructs, the data would not fit the unidimensional

Rasch model. Recent work with multidimensional models may prove fruitful in managing data from complex instruments.

The effects of item order, or interview question order, on responses may produce dependencies in the data that lead to a failure of invariance, as noted above. Measure stability may be sacrificed if responses differ when items are presented in varying orders. For example, asking the question "are you happy with your life" before asking the question "are you happy with your marriage" may produce answers to the second question within the context of the first. Then the response to the second question would be dependent upon the context set by the first question and the two items would not function independently--a violation of an assumption of the Rasch model. Such order effects have been found to be small with achievement test data (ref), and also with attitudinal data (Frantom, Green, & Lam, 2002). Studies of order effects with clinical interview data that relies on separate coding of responses to individual questions are less accessible. One possible solution to the effects of item ordering would be to combine items into "testlets" and treat each set of questions as an item. This is what data analysts do when responses to interdependent questions are used to generate a single score, for example, a stage of development "score." Some questions may be about opinion, others about frequency of behavior, and so the response scales used would be very different. Bode and Wright (1999) note that recent developments in Rasch measurement include how different rating scales can be combined into a single measure.

McDonald (2000) also found item statistics to be sensitive to instructions for survey completion. The effects of context, such as instructions for completion, item grouping, and use of section headings on Rasch statistics have not been thoroughly investigated.

Some problems with measures can be addressed in a Rasch analysis and some cannot be. We can selectively fix misuse of a response scale, identify and delete malfunctioning items, and drop the data of persons who fail to respond appropriately to the task. Other problems cannot readily be fixed, such as failure to define a trait continuum or failure of items or persons to yield a trait from their responses. But, in either case, helpful information can be gained to aid in decision-making, guide scale improvement, and shed light on the validity of the scales constructed.

References

Anderson, L. A., & Zimmerman, M. A. (1993). Patient and physician perceptions of their relationship and patient satisfaction: A study of chronic disease management. *Patient Education and Counseling, 20*, 27-36.

Beck, C. T., & Gable, R. K. (2001). Item response theory in affective instrument development: An illustration. *Journal of Nursing Measurement, 9*, 5-22.

Benson, J. & Clark (1982). A guide for instrument development and validation. *The American Journal of Occupational Therapy, 36*(12), 789-800.

Blumenthal J. (1994) Quality of life and recovery after cardiac surgery. *Psychosomatic Medicine, 56*, 213-215.

Bode, R. K., & Wright, B. D. (1999). Rasch measurement in higher education. In Smart, J. C., & Tierney, W. G. (Eds.), *Higher Education: Handbook of Theory and Research, Volume XIV*. NY: Agathon Press.

Bond, T., & Fox, C. (2001). *Applying the Rasch model*. Mahwah, NJ: Lawrence Erlbaum Associates.

Converse, J. M., & Presser, S. (1986). *Survey questions.* Newbury Park, CA: Sage.

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Newbury Park, CA: Sage.

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology, 45*, 30-45.

Frantom, C. G. (2001). *Paternalism and the myth of perfection: Test and measurement of a theory underlying physicians' perceptions of patient autonomy*. Unpublished doctoral dissertation, University of Denver, Denver, CO.

Frantom, C. G., Green, K. E., & Lam, T. C. M. (2002). Item grouping effects on invariance of attitude items. *Journal of Applied Measurement, 3*, 38-49.

Frost, R.O., Marten, P., Lahart, C., & Rosenblate, R. (1990). The dimensions of perfectionism. *Cognitive Therapy and Research, 14*(5), 449-468.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement,1*, 152-176.

Kauffman, C. L. (1983). Informed consent and patient decision-making: Two decades of research. *Social Science Medicine, 17*, 1657-1664.

Kroger, J., & Green, K. (1994). Factor analytic structure and stability of the Separation-Individuation Test of Adolescence. *Journal of Clinical Psychology, 50*, 772-785.

Levine, J. B., Green, C. J., & Millon, T. (1986). The Separation-Individuation Test of Adolescence. *Journal of Personality Assessment, 50*, 123-137.

Levine, J. B., & Saintonge, S. (1993). Psychometric properties of the Separation-Individuation Test of Adolescence with a clinical population. *Journal of Clinical Psychology, 49*, 492-507.

Linacre, J. M. (2002). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85-106.

Linacre, J. M., & Wright, B. D. (1999-2001). *Winsteps.* University of Chicago: MESA Press.

McDonald, J. (2001*). The susceptibility of item responses to instructions for completion*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April.

Pacht, A. R. (1984). Reflections on perfectionism. *American Psychologist, 39*, 386-390.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Smith, E. V., Lawless, K. A., Curda, L., & Curda, S. (1999). Measuring change in efficacy. *Popular Measurement, 2*, 31-33.

Smith, R. M. (1992). *Application of Rasch measurement*. Chicago: MESA Press.

Smith, R. M., Schumacker, R. E., & Bush, J. M. (1995, April). *Using item mean squares to evaluate fit to the Rasch model.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Snyder, S., & Sheehan, R. (1992). The Rasch measurement model: An introduction. *Journal of Early Intervention, 16*, 87-95.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions*. SF: Jossey-Bass.

Wolfe, E. W., & Chiu, C. W. T. (1999). Measuring change across multiple occasions using the Rasch rating scale model. *Journal of Outcome Measurement, 3*, 360-381.

Wright, B. D., & Stone, M. H. (1979*). Best test design*. Chicago: MESA Press.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.

Table 1. Overall Model Fit Information, Separation, and Mean Logit: Practice Mirroring

```
Summary of 131 Measured Persons
             RAW                            MODEL         INFIT        OUTFIT
           SCORE     COUNT     MEASURE    ERROR       MNSQ   ZSTD   MNSQ   ZSTD
----------------------------------------------------------------------------
MEAN       42.3       15.0        -.36      .36        .99    -.3    .99    -.3
S.D.        7.6         .0         .98      .02        .56    1.5    .55    1.5
MAX.       72.0       15.0        4.18      .62       3.87    4.7   3.85    4.6
MIN.       23.0       15.0       -2.89      .35        .18   -3.4    .18   -3.4
----------------------------------------------------------------------------
 REAL RMSE    .39  ADJ.SD      .90  SEPARATION  2.30  PERSON RELIABILITY  .84
MODEL RMSE    .36  ADJ.SD      .92  SEPARATION  2.56  PERSON RELIABILITY  .87
S.E. OF PERSON MEAN = .09
____________________________________________________________________________

Summary of 15 Measured Items
MEAN      369.1      131.0         .00      .12        .99    -.4    .99    -.4
S.D.       41.0         .0         .59      .00        .38    2.8    .38    2.8
MAX.      434.0      131.0         .95      .12       2.16    7.1   2.15    6.9
MIN.      303.0      131.0        -.94      .12        .53   -4.6    .53   -4.7
-----------------------------------------------------------------------------
 REAL RMSE    .13  ADJ.SD      .58  SEPARATION  4.51   ITEM  RELIABILITY  .95
MODEL RMSE    .12  ADJ.SD      .58  SEPARATION  4.81   ITEM  RELIABILITY  .96
S.E. OF  ITEM MEAN = .16
_____________________________________________________________________________
```

Table 2. Response Scale Use: Practice Mirroring

```
_______________________________________________________________

SUMMARY OF MEASURED STEPS
---------------------------------------------------------------
|CATEGORY OBSVD |OBSVD SAMPLE|INFIT OUTFIT||  STEP    |STEP| THURSTONE|
|SCORE    COUNT |AVRGE EXPECT|  MNSQ  MNSQ||CALIBRATN|S.E.| THRESHOLD|
|--------------------------------------------------------------
|  1       166  | -1.72 -1.67|   .99  1.01||  NONE    |    |
|  2       508  | -1.05  -.99|   .89   .89||   -2.45  |.09 |   -2.66  |
|  3       880  |  -.19  -.26|   .87   .87||   -1.17  |.06 |   -1.07  |
|  4       340  |   .50   .50|  1.01  1.01||    1.06  |.06 |     .99  |
|  5        71  |  1.41  1.72|  1.32  1.30||    2.56  |.14)|    2.74  |
_______________________________________________________________
```

```
        CATEGORY PROBABILITIES: MODES - Step measures at intersections
P      ++------+------+------+------+------+------+------+------++
R  1.0 +                                                         +
O      |                                                         |
B      |                                                         |
A      |                                                         |
B   .8 +11                                                      5+
I      |  11                                                  55 |
L      |    1                                               55   |
I      |     11                                            5     |
T   .6 +       1                  3333                    5      +
Y      |        1               33    333               55       |
    .5 +         1    222     33         33   4444444  5         +
O      |          1*22   22233             344       4*          |
F   .4 +         22 1      322            443        5 44        +
       |       22    1    3   22        44   33    55    44      |
R      |     22       1 33      2      4       3  5        44    |
E      |   22          *1        22  44         3*           44  |
S   .2 + 22          33  1         **          55 3            44+
P      |2          33     11     44  22      55    33            |
O      |         33         11 44      222 55        333         |
N      |    33333         4444*1111   5555*222          33333    |
S   .0 +******************555555555***11111111*******************+
E      ++------+------+------+------+------+------+------+------++
       -4     -3     -2     -1      0      1      2      3      4
                       PERSON [MINUS]  ITEM MEASURE
```

Figure 1. Step Use by Person-Item Measure: Practice Mirroring

Table 3. Item Fit Statistics in Order by Misfit: Practice Mirroring

```
+------------------------------------------------------------------------------------+
|ENTRY    RAW                                   |   INFIT   |  OUTFIT   |SCORE|                  |
|NUMBER  SCORE   COUNT  MEASURE   ERROR|MNSQ  ZSTD|MNSQ  ZSTD|CORR.|  ITEMS           |
|-------------------------------------+----------+----------+-----+------------------|
|    87    434    131    -.94     .12|2.16  7.1 |2.15   6.9|A .36|physical appear  |
|    18    419    131    -.72     .12|1.24   1.8|1.29   2.1|B .41|act like leader  |
|    32    423    131    -.78     .12|1.29   2.1|1.29   2.1|C .42|better off       |
|    13    303    131     .95     .12|1.12   1.0|1.10    .9|D .57|enjoy body       |
|    36    418    131    -.70     .12|1.10    .8|1.09    .7|E .47|amaze myself     |
|    71    357    131     .18     .12|1.07    .6|1.05    .4|F .56|people amazed    |
|    91    364    131     .08     .12|1.01    .1|1.01    .1|G .59|feel special     |
|    65    323    131     .66     .12| .89  -1.0| .88  -1.0|H .75|can tell admired |
|    94    418    131    -.70     .12| .83  -1.5| .83  -1.5|g .51|positive vibes   |
|     2    335    131     .49     .12| .81  -1.7| .80  -1.7|f .65|feel powerful     |
|    92    358    131     .16     .12| .75  -2.2| .76  -2.1|e .56|centre of atten  |
|    83    331    131     .55     .12| .73  -2.5| .72  -2.5|d .72|others impressed |
|    43    358    131     .16     .12| .67  -3.1| .67  -3.1|c .67|people admire me |
|    88    333    131     .52     .12| .65  -3.4| .65  -3.4|b .72|others admire me |
|    97    363    131     .09     .12| .53  -4.6| .53  -4.7|a .75|impressed        |
|-------------------------------------+----------+----------+-----+------------------|
| MEAN    369.   131.     .00     .12| .99   -.4| .99   -.4|     |                  |
| S.D.     41.     0.     .59     .00| .38   2.8| .38   2.8|     |                  |
+------------------------------------------------------------------------------------+
```

Table 4. Fit of Person Responses to Items.

```
TABLE OF POORLY FITTING PERSONS   ( ITEMS IN ENTRY ORDER)
NUMBER - NAME -- POSITION ------ MEASURE - INFIT (MNSQ) OUTFIT

    54  054                            -.54     3.9   A    3.9

     ITEMS:           2 13 18 32 36 43 65 71 83 87 88 91 92 94 97
  RESPONSE:      1:   1  1  3  5  3  3  2  3  5  1  1  4  5  3  1
Z-RESIDUAL:                    2              3 -3        3    -2
```

```
PERSONS                 ITEMS
LOGIT                |         ↑ Higher levels of narcissism
    5                +
                     |
                     |
                     |
                     |
                  X  |
    4                +
                     |
                     |
                     |
                     |
                     |
    3                +
                     |
                     |
                  X  |
                     |
                     |
    2                +
                     |
                    T|
                  X  |
                 XX  |
                  X  |T
    1                +  13-enjoy looking at my body
             XXXXXX  |
          XXXXXXXXX S|S 65-can tell am admired
              XXXXX  |  2-feel powerful  83-others impressed by me  88-others admire
             XXXXXX  |
      XXXXXXXXXXXXX  |  43-admire me  71-amazed by me 92-attention 97-impressed
    0       XXXXXXX  +M 91-feel special
             XXXXXX  |
       XXXXXXXXXXXX M|
         XXXXXXXXXX  |
                 XX  |S 18-act like leader  36-amaze myself   94-get positive vibes
      XXXXXXXXXXXXX  |  32-better off than others
   -1         XXXXX  +  87-knowing physical appearance attractive pleases me
              XXXXX  |T
          XXXXXXXXX S|
               XXXX  |
               XXXX  |
               XXXX  |
   -2             X  +
                  X  |
                 XX T|
                     |
                     |
                  X  |
   -3                +
                     |         ↓ Lower levels of narcissism
```

Figure 2. Item-Person Map: Practicing Mirroring

Table 5. Results of Expert Review of Item Content (10 initial items)

Item	Pat.	Per.	Both	Neith	Patient Autonomy	
					Yes	No
1. Physicians should not make mistakes.	0	5	1	0	2	4
2. Patients believe that physicians should not make mistakes.	0	4	1	1	2	4
3. Physicians should not tell patients when they are uncertain in medical matters.	**4**	**1**	**1**	**0**	**5**	**1**
4. It is not a physician's responsibility to know all of the answers to medical questions.	**0**	**5**	**0**	**1**	**3**	**3**
5. It is important to treatment outcome that patients see physicians as being certain.	2	1	3	0	3	2
6. Patients are generally capable of managing their own health care.	2	0	0	4	5	1
7. What patients don't know won't hurt them.	**6**	**0**	**0**	**0**	**6**	**0**
8. Patients assume that doctors know more than patients do.	**4**	**2**	**0**	**0**	**5**	**1**
9. Physicians usually make the right decisions.	**1**	**1**	**4**	**0**	**4**	**2**
10. It is better for the patient if the physician does not disclose a mistake.	**1**	**1**	**4**	**0**	**6**	**0**

Note: Items in bold-type were retained for the pilot study.

Table 6. Overall Model Fit Information for 'Physician Disclosure' Subscale
(158 persons, 9 items)

	Infit		Outfit				
	MNSQ	ZSTD	MNSQ	ZSTD	Adj.SD	Separation	Reliability
Persons					1.71	1.73	.75
Mean	1.00	-.3	.98	-.3			
S.D.	.70	1.4	.73	1.4			
Items					.83	5.23	.96
Mean	.99	-.3	.98	-.3			
S.D.	.25	2.3	.27	2.2			

Table 7. Overall Model Fit Information for 'Patient Needs' Subscale
(160 persons, 6 items)

	Infit		Outfit					
	MNSQ	ZSTD	MNSQ	ZSTD	Adj.SD	Separation	Reliability	SE Mean
Persons					1.09	1.30	.63	.11
Mean	1.00	-.3	.99	-.3				
S.D.	.80	1.3	.77	1.2				
Items					1.23	9.04	.99	.55
Mean	.99	-.2	.99	-.1				
S.D.	.13	1.2	.13	1.2				

Table 8. Item Fit Statistics in Misfit Order: Physician Disclosure

Entry			Infit		Oufit		Point Biserial
Number	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD	Correlation
4	-.07	.16	1.42	3.0	1.40	2.8	.39
15	-1.19	.14	1.19	1.6	1.25	2.0	.39
20	.63	.16	1.18	1.4	1.11	.8	.37
11	-1.20	.14	.99	-.1	1.05	.4	.47
9	-.26	.15	.99	0.0	1.01	.1	.40
16	-.67	.15	.93	-.6	.95	-.4	.52
5	.73	.16	.90	-.9	.86	-1.1	.54
6	1.29	.16	.83	-1.5	.79	-1.6	.54
14	.73	.16	.46	-5.6	.42	-5.4	.60
Mean	0.00	.15	.99	-.3	.98	-.3	
SD	.85	.01	.25	2.3	.27	2.2	

Table 9. Item Fit Statistics in Misfit Order: Patient Needs

Entry			Infit		Oufit		Point Biserial
Number	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD	Correlation
8	450	.13	1.16	1.5	1.17	1.5	.42
2	422	.13	1.13	1.2	1.15	1.4	.34
17	465	.13	.95	-.4	.99	-.1	.25
12	274	.14	.95	-.4	.92	-.6	.34
3	362	.13	.92	-.7	.88	-1.1	.49
18	382	.13	.79	-2.1	.81	-1.8	.47
Mean	392	.13	.99	-.2	.99	-.1	
SD	64	.01	.13	1.2	.13	1.2	

```
R  1.0 _______________________________________________________________
O      |                                                             |
B      |                                                             |
A      |                                                          444|
B   .8 +11                 22222                                44   +
I      |  1              222     22                            4     |
L      |   11         22          22                         44      |
I      |     1       2              22         33333        4        |
T   .6 +      1    22                 2     333     33     4         +
Y      |       11 2                    2   3          33  4          |
    .5 +         *                      233             34           +
O      |       22 1                     322             43           |
F   .4 +      2    1                   3   2          44  33         +
       |     2      1                33     2        4      3        |
R      |   22        11             3        2      4        33      |
E      |  2           1          33          22   4           3      |
S   .2 +22              11       3              244             33   +
P      |                  11  333              4422               333|
O      |                   3**1              44    222               |
N      |              33333    111111   44444         2222           |
S   .0 +**************444444444444444***111111111111111111***********+
       -5    -4    -3    -2    -1     0     1     2     3     4     5
                         Person (Minus) Item Measure
      Summary of Measured Steps:
      Step Label                  1          2          3                 4
      Average Measure          -3.56        -1.50        .05            .86
```

Figure 3. Step Use by Person-Item Measure: 'Physician Disclosure' Subscale

```
SD  Person Ability  Item Agreeability
    2                +  ↑ Harder to Agree With
                  .  |
                     |Q
                  # Q|
                  #  |      6 It is best to protect patients from information that will upset them. (2) D
    1                +
                 .#  |S    14 It does no good to disclose a medical error to patients.(1) D
                     |      5 It would be harmful to patients if physicians disclosed mistakes. (1) D
                .##  |     20 Physicians should not tell patients when they are uncertain in medical matters. (2) M/D
                .##  |
    0               S+M     4 It is right for physicians to make decisions for their patients. (2) M
           ########  |      9 Disclosing physician mistakes will undermine patient trust. (1) M
                   |
           ########  |     16 Admitting a mistake to a patient is not important in most medical situations. (1) M
                     |S
   -1                +
      .############  | 11/15 It is sometimes in the best interests of the pt. for physicians not to admit mistakes.(2)M
                     |       /Physicians are in the best position to make choices for patients about treatment. (2)M
       .############ M|Q
                     |
   -2       #######  +
                     |
             .#####  |
                     |
                     |
   -3         .####  +
                    S|
                 ##  |
                     |
               .###  |
   -4                +
                     |
                .##  |
                    Q|
                     |
   -5            ##  +
                     |
                     |
                     |
                     | ↓ Easier to Agree With
   -6           ###  +
M = Mean                         (1) = Perfectionism Item
S = 1 Standard Deviation         (2) = Paternalism Item
Q = 2 Standard Deviations         #  =  Logit position for several persons
M = Moderate, D = Difficult: Expert reviewer ratings of item difficulty level
```

Figure 4. Item-Person Map: Physician Disclosure

```
SD  Person Ability  Item Agreeability
    3                +
                     | ↑ Harder to Agree With
                 ## Q|
                     |Q 12  Information about health care options just confuses patients. (2) M/D
                .##  |
    2                +
                     |
               .###  |
                     |
                    S|S
    1         .####  +
                     |
           .#######  |  3  Most patients only want to know the best option. (2) M
                     |
           .#######  |  18  Most patients want physicians to make decisions for them.  (2) M
    0               M+M
                     |
         .#########  |
                     |  2  Patients are easily overwhelmed when given all their options.  (2) M
                     |
   -1      .#######  +  8  Patients want physicians to have perfect knowledge.  (1)  Added
                     |S
               .### S|  17  Most physicians feel they need to always have an answer to a patient's question. (1)
Added
                     |
                     |
   -2                +
               .###  |
                     |Q
                     |
                 .# Q|
   -3                +
                     |
                  .  |
                     |
                     | ↓ Easier to Agree With
   -4                +
M = Mean                         (1)  = Perfectionism Item
S = 1 Standard Deviation         (2)  = Paternalism Item
Q = 2 Standard Deviations         #   = Logit position for several persons
M  = Moderate, D = Difficult: Expert reviewer ratings of item difficulty level
```

Figure 5. Item-Person Map: Patient Needs