Open Science Journal of Statistics and Application
2014; 2(1): 17-23
Published online March 10, 2014 (http://www.openscienceonline.com/journal/osjsa)

Item response theory in healthcare research: Reasons for extending common models

P. J. Veazie

Department of Public Health Sciences, University of Rochester, Rochester, USA

Email Address

peter_veazie@urmc.rochester.edu

To cite this article

P. J. Veazie. Item Response Theory in Health Services Research: Reasons for Extending Common Models, *Open Science Journal of Statistics and Application.* Vol. 2, No. 1, 2014, pp. 17-23

Abstract

Item Response Theory (IRT) has enjoyed increased interest in recent years as a method for scaling health-related constructs. However, the context of this application is different from the context of IRT's development. The shift in context has important implications for proper model specification. This paper reviews the common one, two, and three parameter IRT models, discusses their limitations as used in healthcare research, and argues for the four parameter hierarchical model as an alternative. The use of IRT as a pragmatic means to test item bias has merit, but the general use of IRT in healthcare research without consideration of underlying assumptions may lead to a less appropriate application. Healthcare research would benefit from development of models such as the 4-parameter IRT to account for plausible underlying assumptions.

Keywords

Item Response Theory, Health Measurement Scales

1. Introduction

In health research, Item Response Theory (IRT) has gained advocacy in recent years as a method for constructing, combining, and assessing health-related scales; for example, much of a 2000 issue of *Medical Care* was dedicated to articles on IRT (Cella and Chang 2000; Hambleton 2000; Hays, Morales, and Reise 2000; McHorney and Cohen 2000; Ware, Bjorner, and Kosinski 2000). IRT is often used to measure unobservable constructs such as health status and quality of life. This use departs from IRT's origin in test development—a departure with important consequences. This paper motivates extending the current IRT models to better represent phenomena of interest to healthcare researchers. A critique of common models is given and an argument for their extension provided. The goal is to spur further interest in creating methods for estimation and testing of extended models.

This paper addresses scale development in a context presupposing a latent quality that generates variation in manifest variables as opposed to a context in which manifest variables are combined to form an index. The former is the domain of psychometric techniques such as classical scaling, factor analysis and IRT (Nunnally and Bernstein 1994); the latter is the domain of Clinimetrics (de Vet, Terwee, and Bouter 2003a, 2003b; Fava and Belaise 2005; Fayers and Hand 2002; Feinstein 1982, 1983, 1987; Marx et al. 1999; Streiner 2003). Feinstein (Feinstein 1987), the father of the clinimetric terminology, defines the clinimetrics as "arbitrary ratings, scales indexes, instruments or other expressions that have been created as "measurements" for those clinical phenomena that cannot be measured in the customary dimensions of laboratory data."; what Fayers and Hand (Fayers and Hand 2002) describe as choosing and emphasizing "…the most important attributes to be included in the index, using multiple items which are not expected to be homogeneous because they indicate different aspects of a complex clinical phenomenon." See for examples the Apgar scale for newborns (Feinstein 1999) or the Jones criteria for rheumatic fever (Feinstein 1982, 1983). If one seeks to index a heterogeneous complex such as clinical syndromes

or indicators such as socio-economic status (and arguably quality of life measures as well), then a clinimetric approach may be advisable over a psychometric one that assumes underlying latent qualities. The former method does not require correlation among the manifest variables, whereas the latter depends on it. Notwithstanding the merits of a clinimetric approach, this paper focuses solely on a critique of IRT. Moreover, the focus is on models of dichotomous responses in which the probability of response is monotonically increasing with the latent trait; unfolding response models (e.g. (Johnson and Junker 2003)), in which monotonicity is not imposed, are not considered here.

When IRT is used to represent the correspondence between a latent ability and the probability of correctly performing a specific test of that ability, the test is a direct manifestation of the underlying attribute. For example, endorsing as true the statement "1 + 1 = 2" directly tests the ability of an individual to add numbers (assuming unidimensionality). Greater ability corresponds to higher probability of a correct answer if the test subject intends to give the correct answer. The correct answer is invariant over the scale of abilities: one plus one equals two regardless of the ability of the test subject. Similarly, a test of reading comprehension would entail performing a task that requires reading comprehension. For example, the statement, "the first paragraph of this paper addresses the use of IRT in healthcare research" would measure reading comprehension. The set of possible answers {true, false} contains the correct answer regardless of the test subject's ability. Even if the test question were open ended such as "what is the topic of the first paragraph?" there would be a correct answer (or set of correct answers) and a corresponding set of incorrect answers (possibly only defined as *not correct* and therefore not explicitly listed).

In general, a test of ability requires an exhibition of that ability in conjunction with the assumption that the test subject engages the ability. This is often not the situation in health research. In health research IRT is often used to measure an underlying construct based on responses being correlated with the construct but not direct tests of the construct. Examples include measuring quality of life (Leplege et al. 1997), severity of illness such as asthma (Morris et al. 1996), severity of effects such as pain (Kopec et al. 1996; Tesio, Granger, and Fiedler 1997), and health behaviors such as drug and alcohol abuse (Kirisci, Tarter, and Hsu 1994; Muthen 1996).

Quality of life is supposed to generate a positive response to the statement "I generally enjoy what I do" (Leplege et al. 1997). Notice, however, that because quality of life is not defined by the act of endorsing that statement, the response is not a test of quality of life: it has no independently correct answer. Here the subject is not assumed to intend answering correctly, only honestly. Indeed, if the subjects were intending to answer according to some perceived notion of correctness, they would bias the measurement to the extent that they were not answering honestly. This weakening of the assumption of intended correctness adds an element of uncertainty to the measure at any given level of the underlying trait. Where previously we knew whether a given answer was correct, now we can only assume the answer was honestly provided: both endorsement and non-endorsement can be proper for each individual at a given trait level. This is where the use of IRT in healthcare research often departs from its use in testing; a departure with implications for model specification.

2. Parameterization

Common parameterizations in healthcare research of item characteristic curves using the logistic model are given in the first 3 rows of Table 1 along with corresponding graphs. The 1-parameter model allows each item to differ in its discrimination parameter (a_j), thereby differently discriminating the underlying trait. The 2-parameter model also allows items to differ in location (d_j), thereby allowing each item to be sensitive to different parts of the underlying trait scale. The classic 3-parameter model allows for different lower limits to the probability of endorsing the item as the trait level decreases. The 3-parameter model was an important step in representing ability tests for which the answer can be guessed; it explicitly recognized that is such cases even a person with no ability has a nonzero probability of endorsing the correct answer if they guess. The classic 3-parameter model assumes that as the trait increases, the probability of endorsing the item approaches 1.

Under the assumption that a subject intends to answer correctly and the fact that success is directly a function of ability, it is reasonable to assume that the probability of a correct response on a given test is monotonically nondecreasing and approaches 1 as ability increases. In other words, for any level of test difficulty there is a level of ability such that the probability of correctly performing the test is arbitrarily close to 1. This requires the modeling assumption that ability can in principle increase without bound. It could be argued that because the test must be conceptualized and identified with a correct answer set, the very fact that a test question is posed presupposes an extant minimally sufficient corresponding ability; otherwise a correct answer would not be identified.

When the underlying trait is not directly measured by the test and any answer is appropriate for honestly responding subjects, it is less plausible to assume that the upper limit of the probability converges to 1. It is not necessary that there exist a level of the underlying trait at which the probability of endorsing a statement in a survey need approach 1 arbitrarily close. This suggests that the IRT models for these measurement tasks based on dichotomous responses ought to include an upper limit parameter. For example, as shown in row 4 of Table 1, using the Logit

model the specification would be

$$P(y_{i,j}=1)=L_j+\frac{U_j-L_j}{1+e^{-a_j\cdot(\tau_i-d_j)}}$$

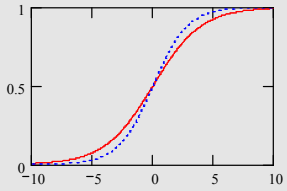
where L_j is the probability observation i endorses item j as trait $\tau \to -\infty$ (the trait goes infinitely low); U_j is the probability of endorsing the item as $\tau \to \infty$ (the trait goes infinitely high); and a and d are the scale and location parameters of the item (discrimination and "difficulty" respectively). The corresponding figure given in Table 1 shows two specifications: One, represented by the solid line, corresponds to a 2-parameter model with parameters a = 0.5, d = 0, L = 0, and U = 1; the other, represented by the dashed line, has parameters a = 0.5, d = 0, L = 0.2, and U = 0.8. Not only are the limiting probabilities different, but with the same discrimination parameter a, the actual discrimination is influenced by the effect of the difference in upper and lower limits on the slope:

$$\frac{dP_{ij}}{d\tau_i}=(U_j-L_j)\cdot a_j\cdot P_{ij}\cdot(1-P_{ij})$$

Table 1. *Logistic item characteristic curves for 1, 2, 3, and 4 parameter models. The 1-parameter model allows different discrimination parameters (a_j); the 2-parameter model allows different discrimination and intercepts (d_j); the 3-parameter model allows different discrimination, intercepts, and lower limits (L_j); the 4-parameter model allows different discrimination, intercepts, lower limits, and upper limits (U_j).*

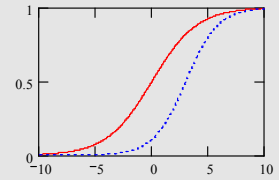
1-parameter model:

$$P(y_{i,j}=1\mid\tau)=\frac{1}{1+e^{-a_j\cdot(\tau_i)}}$$



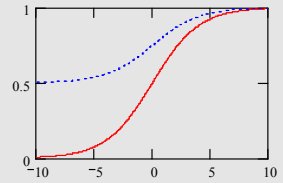
2-parameter model:

$$P(y_{i,j}=1\mid\tau)=\frac{1}{1+e^{-a_j\cdot(\tau_i-d_j)}}$$



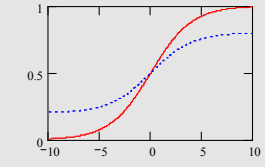
3-parameter model:

$$P(y_{i,j}=1\mid\tau)=L_j+\frac{1-L_j}{1+e^{-a_j\cdot(\tau_i-d_j)}}$$



4-parameter model:

$$P(y_{i,j}=1\mid\tau)=L_j+\frac{U_j-L_j}{1+e^{-a_j\cdot(\tau_i-d_j)}}$$



Setting L = 0 and U =1 in this equation yields the slope for the usual Logit IRT model.

Simplifications available for some tests of ability are also not as plausible in the health care context. For a test of ability, individuals with no ability can randomly guess at the correct answer; if all items on the test have an equal number of possible answers, then the probability of a correct guess is the same across items and $L_j = L$ for all items j. The figure in Table 1 associated with the 3-parameter model shows a graph with the lower limit probability equal to 0.5 associated with random guessing the correct answer among a dichotomous choice set. In the present context, however, assuming a common probability of endorsing all items when a person's latent trait is infinitely negative is not plausible (although statistically testable in the model).

Methods papers written to advocate the use of IRT in healthcare research tend to stay with 1, 2, or 3 parameter models (Revicki and Cella 1997; van Alphen et al. 1994). Although the 4-parameter alternative may require a fairly

large sample size and considerable computation time, it is proposed here that the 4-parameter model can better represent some phenomena of interest in healthcare research, and given the continual improvement in computer computational capabilities, development of estimation and testing methods for the 4-parameter model now seems warranted. By accepting a 3-parameter model, one has accepted the need to properly model the lower limit of the item characteristic curve. The argument for the 4-parameter model follow a similar logic, only applied to the upper limit of the item characteristic curve.

The key assumptions that differentiate the two, three, and four parameter models are those regarding the limiting probabilities of endorsing an item as the latent trait or quality decreases and increases. The 2-parameter model fixes the lower limit to 0 and the upper limit to 1; this implies that those at one end of the scale will certainly not endorse the item whereas those at the other end of the scale will most certainly endorse the item. The classic 3-parameter model fixes the upper limit to 1 but allows the lower limit to be freely estimated; this implies that those at one end of the scale will always have some non-zero probability of endorsing the item. The 4-parameter model allows both the lower and upper limits to be free, implying at both ends of the scale responses are less than certain. Clearly the 2 and 3-parameter models are merely special cases of the 4-parameter model. As such, the 4-parameter model represents a more general monotonic item characteristic curve of which the other models are special cases. Although this paper focuses on the 4-parameter model, it should be noted that an alternative 3-parameter model may be appropriate in certain circumstances as well. Specifically, if the lower limit is fixed to 0 but the upper limit is free, then we have a 3-parameter model that implies a certainty of response for the lower limit of the underlying trait or quality but a less than certain response in the upper limit.

Making appropriate a priori assumptions regarding these models will depend on the nature of the problem being studied. Consider for example a study by Landrum et. al. (Landrum, Bronskill, and Normand 2000) which include in their analysis of hospital quality an indicator of coronary angiography (CA) among patients for which angiography is indicated as a necessary procedure by consensus guideline criteria. Hospital quality is considered a latent trait that generates a probability of giving CA to patients for whom it is indicated as necessary. They use a 2-parameter item characteristic curve, thereby assuming the probability of providing CA approaches 0 as quality becomes decreasingly low, and the probability of providing CA approaches 1 as quality becomes increasingly high. Alternatively, however, we might consider that consensus guidelines are not sufficiently subtle to identify the specific needs of each individual and some patients who meet the guideline criteria for needing CA can be properly judged as not a candidate for CA by a hospital of the highest quality. In this case, we may suppose that the probability of providing CA to guideline-identified patients is less than 1 even for the highest quality of hospital care. Hence, a 3-parameter model in which the lower limit is fixed to 0 and the upper limit is free may be a better selection. Moreover, hospitals may give CA to some patients that have counter indications not unambiguously identified in the consensus guidelines and therefore the lowest possible quality hospitals may still have a non-zero probability of giving CA. In this case, a 4-parameter model that allows both upper and lower limits to be free may be appropriate. The a priori selection of a model depends on how the problem is conceptualized and what the observed variables represent. Clearly the development of statistical tests to empirically adjudicate model specification would be advantageous in this case.

3. Parameter Heterogeneity

Another departure from the traditional IRT development is that a test of ability (e.g. the ability to add numbers) provides for a level of independence not available in health measurement of constructs such as quality of life. Each observation from a population with a given ability level necessarily has the same probability of a correct answer (assuming unidimensionality); if the probabilities are different, then the ability must also be different. In the case of measuring quality of life, this independence does not hold. Each individual will likely have a different characteristic function associated with each item due to the influence of their life experience on their interpretation of the statement and the threshold of the trait corresponding to endorsing the statement.

The standard IRT model is expressed as

$$P(y_{ij} = 1 | \beta_j, \theta_i) = g(\beta_j, \theta_i)$$

where $y_{ij} = 1$ indicates observation i correctly endorses item j; β_j is a vector of parameters associated with the test item j; and θ_i is the underlying ability of person i. The only variation associated with individuals in this model is associated with variation in ability θ_i. It is assumed that β_j is invariant across individuals. The formulation for constructs such as quality of life is better represented as

$$P(y_{ij} = 1 | \beta_{ij}, \theta_i) = g(\beta_{ij}, \theta_i)$$

{{REMOVE INDENT}}in which the parameters of the distribution are a function of the individual. This lack of independence is not a function of additional latent traits (i.e. it is not an issue of multidimensionality). Rather than a function of differential abilities on unaccounted dimensions, it is a function of individuality. Each person may have the same ability to read and understand the sentence: however, they have a different correspondence between the mutual understanding of the English and the level of the underlying construct. Note that where English can be commonly understood among people, the correspondences

with internal constructs that are not directly measured are less likely to generate a mutual understanding because they are not ostensibly presented for common learning.

To the extent that the variation in β is explained by categories in a population, the model is said to exhibit differential item functioning. However, statistically identifying differential item functioning requires the prior identification of an appropriate categorization; for example, gender race, culture, etc. And, when differential item functioning is identified for some partition of the population, the within group variation of the item characteristics is assumed invariant. This is an assumption that is not necessarily plausible when all answers are deemed "correct" if the reply to the test statement is honest.

Under these circumstances the item parameters (i.e. discrimination, difficulty, upper and lower limits) should be treated as individual-specific parameters (i.e., a vector β_{ij}) and estimation focused on appropriate summary statistics of the distribution of those parameters—for example, mean values:

$$P(y_{ij} = 1| \beta_{ij}, \theta_i) = g(\beta_{ij}, \theta_i),$$
$$\beta_{ij} \sim F(\beta_j).$$

In this case the parameter vector β_j represents the parameters of the distribution F describing the individual-specific parameters β_{ij}. An appropriate model to account for this non-categorical variation in the item parameters is a hierarchical modeling scheme whereby conditional distributions of the parameters can be specified as functions of specific individual characteristics. This may be addressed through the use of random coefficient models or Bayesian methods (Albert 1992; Ghosh et al. 2000; Landrum et al. 2000; Sahu 1998; Tsutakawa and Lin 1986).

4. Pragmatic Utility

Though a good theoretical motivation is desirable, one is not always necessary. One set of studies that used model fit as their criteria for assessment entailed the comparison between Likert and Rasch scoring methods for a common physical functioning scale (McHorney, Haley, and Ware 1997; Raczek et al. 1998). The authors argue for the use of the Rasch method due to its superior performance, though the findings were less definite in the McHorney study. Another useful application of IRT in this body of research is the testing of some specific questions regarding the nature of a preexisting scale; specifically, the use of IRT to investigate item bias. Typically, these research projects are focused on determining if a particular existing instrument manifests differential item bias across specified groups of interest. The importance of these inquiries is to assure that these instruments can provide for consistent conclusions. For example, measures used to determine dementia have been criticized for item bias across racial and educational groupings. Teresi et al. (1995) investigated the merit of these assertions by using IRT models to test for differential item bias. The focus of their study was not the development of a new scale based on IRT methods of scoring; instead, they were specifically addressing the hypothesis of DIF without impugning the scoring methodology of the overall test. They found a number of items that exhibited item bias across groups based on race and based on education level. It is interesting to note that they tested the usual one, two and three parameter models, but did not apparently consider a four-parameter model as an option.

A second example extends the usefulness of IRT to language translation equivalence. The SF-36, a common health status survey, has been translated into multiple languages. This creates the problem of achieving translation equivalence. In order for the instrument to operate in the same fashion across languages and cultures, the items must be translated in a way such that its association with the underlying construct remains invariant up to the transformation in the scale units. Bjorner et al. (1998) evaluated translation equivalence using IRT to investigate differential item functioning across different countries where the survey was translated into the appropriate language. Their assertion is that if an item is functioning in a similar fashion across cultures, then its translation must be appropriate. Conversely, if DIF is found, then the translation is not appropriate and should be modified. This test of translation equivalence automatically accounts for context and culture; it is not simply a test for correct word translations and corresponding grammatical consistency.

These are examples of appropriate uses of classic IRT models in healthcare research in the absence of a serious theoretical base. However, one could object this leniency is contrary to the lack of charity in the first part of the paper. Do not those concerns apply to the DIF examples just presented? Yes, but they are not as damaging. Even if the item parameters vary within groups, the estimates based on groups can be considered the mean values of the groups' distributions of item parameter values. An IRT comparison between groups for DIF is then testing if the item functions on average differently between groups. Nonetheless, applying a 4-parameter random-coefficient model in these applications may prove useful as well.

5. Conclusion

IRT is a theory that can benefit healthcare research, but the extent and limitations of its usefulness has yet to be fully explored. The use of IRT as a pragmatic means to test item bias has merit; however, the general use of IRT in healthcare research without consideration of underlying assumptions may lead to a less appropriate application. This failing has important implications for model specification. Healthcare research would benefit from the investigation of extended models such as the 4-parameter models as discussed above to determine whether the proposed extensions provide a useful addition to ~~HSR~~ healthcare research methods. Specifically, healthcare

research would benefit greatly from efforts to develop feasible and efficient estimators, determine model identification criteria, and provide model specification tests (e.g., determination of unidimensionality and the need for the heterogeneity in item parameters). At a minimum, such investigations will provide insight into the empirical consequences of not extending the current models when theory suggests otherwise: it may turn out that ignoring the theoretically implied extensions comes at little empirical cost, or specific circumstances may be determined where such simplifying assumptions are too costly to ignore. In either case, the methods of healthcare research will benefit from researchers extending the currently used IRT models to account for plausible underlying assumptions.

References

[1] Albert, J. H. 1992. "Bayesian estimation of normal Ogive item response curves using Gibbs sampling." *Journal of Educational Statistics* 17: 251-69.

[2] Bjorner, J. B., S. Kreiner, J. E. Ware, M. T. Damsgaard, and P. Bech. 1998. "Differential Item Functioning in the Danish Translation of the SF-36." *Journal of Clinical Epidemiology* 51(11): 1189-202.

[3] Cella, D. P. and C.-H. P. Chang. 2000. "A Discussion of Item Response Theory and Its Applications in Health Status Assessment." *Medical Care* 38(9 Suppl. II): 66-72.

[4] de Vet, H. C. W., C. B. Terwee, and L. M. Bouter. 2003a. "Clinimetrics and psychometrics: two sides of the same coin." *Journal of Clincial Epidemiology* 56: 1146-47.

[5] de Vet, H. C. W., C. B. Terwee, and L. M. Bouter. 2003b. "Current challenges in clinimetrics." *Journal of Clinical Epidemiology* 56: 1137-41.

[6] Fava, G. A. and C. Belaise. 2005. "A discussion on the role of clinimetrics and the misleading effects of psychometric theory." *Journal of Clinical Epidemiology* 58(8): 753-56.

[7] Fayers, P. M. and D. J. Hand. 2002. "Causal variables, indicator variables and measurement scales: an example from quality of life." *Journal of the Royal Statistical Society A* 165(2): 233-61.

[8] Feinstein, A. R. 1982. "The Jones criteria and the challenge of clinimetrics." *Circulation* 66: 1-5.

[9] Feinstein, A. R. 1983. "An additional science for clinical medicine: The development of clinimetrics." *Annals of Internal Medicine* 99: 843-48.

[10] Feinstein, A. R. 1987. *Clinimetrics*. New Haven: Yale University Press.

[11] Feinstein, A. R. 1999. "Multi-item "Instruments" vs Virginia Apgar's Principles of Clinimetrics." *Archives of Internal Medicine* 159(2): 125-28.

[12] Ghosh, M., A. Ghosh, M.-H. Chen, and A. Agresti. 2000. "Noninformative priors for one-parameter item response models." *Journal of Statistical Planning and Inference* 88: 99-115.

[13] Hambleton, R. K. P. 2000. "Emergence of Item Response Modeling in Instrument Development and Data Analysis." *Medical Care* 38(9 Suppl. II): 60-65.

[14] Hays, R. D. P., L. S. M. D. M. P. H. Morales, and S. P. P. Reise. 2000. "Item Response Theory and Health Outcomes Measurement in the 21st Century." *Medical Care* 38(9 Suppl. II): 28-42.

[15] Johnson, M. S. and B. W. Junker. 2003. "Using data augmentation and Markov Chain Monte Carlo for the estimation of unfolding response models." *Journal of Educational and Behavioral Statistics* 28(3): 195-230.

[16] Kirisci, L., R. E. Tarter, and T. Hsu. 1994. "Fitting a two-parameter logistic item response model to clarify the psychometric properties of the drug use screening inventory for adolescent alcohol and drug abusers." *Alcoholism: Clinical and Experimental Research* 18(6): 1335-41.

[17] Kopec, J. A., J. M. Esdaile, M. Abrahamowicz, L. Abenhaim, S. Wood-Dauphinee, D. L. Lamping, and J. I. Williams. 1996. "The Quebec back pain disability scale: conceptualization and development." *Journal of Clinical Epidemiology* 49(2): 151-61.

[18] Landrum, M. B., S. E. Bronskill, and S.-L. T. Normand. 2000. "Analytic methods for constructing cross-sectional profiles of health care providers." *Health Services and Outcomes Research Methodology* 1(1): 23-47.

[19] Leplege, A., N. Rude, E. Ecosse, R. Ceinos, E. Dohin, and J. Pouchot. 1997. "Measuring quality of life from the point of view of HIV-positive subjects: the HIV-QL31." *Quality of Life Research* 6: 585-94.

[20] Marx, R. G., C. Bombardier, S. Hogg-Johnson, and J. G. Wright. 1999. "Clinimetric and Psychometric Strategies for Development of a Health Measurement Scale." *Journal of Clinical Epidemiology* 52(2): 105-11.

[21] McHorney, C. A., S. M. Haley, and J. E. Ware. 1997. "Evaluation of the MOS SF-36 Physician Functioning Scale (PF-10): II. Comparison of Relative precision using Likert and Rasch scoring methods." *Journal of Clinical Epidemiology* 50(4): 451-61.

[22] McHorney, C. A. P. and A. S. P. Cohen. 2000. "Equating Health Status Measures With Item Response Theory: Illustrations With Functional Status Items." *Medical Care* 38(9 Suppl. II): 43-59.

[23] Morris, M. V., M. J. Abramson, M. J. Rosieer, and R. P. Strasser. 1996. "Assessment of the severity of Asthma in a family practice." *Journal of Asthma* 33(6): 425-36.

[24] Muthen, B. O. 1996. "Psychometric evaluation of diagnostic criteria: application to a two-dimensional model of alcohol abuse and dependence." *Drug and Alcohol Dependence* 41: 101-12.

[25] Nunnally, J. C. and I. H. Bernstein. 1994. *Psychometric Theory*. New York: McGraw-Hill.

[26] Raczek, A. E., J. E. Ware, J. B. Bjorner, B. Gandek, S. M. Haley, N. K. Aaronson, G. Apolone, P. Bech, J. E. Brazier, M. Bullinger, and M. Sullivan. 1998. "Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA project." *Journal of Clinical Epidemiology* 51(11): 1203-14.

[27] Revicki, D. A. and D. F. Cella. 1997. "Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing." *Quality of Life Research* 6: 595-600.

[28] Sahu, S. K. 1998. "Bayesian estimation and model choice in item response models." School of Mathematics: University of Wales, CA.

[29] Streiner, D. L. 2003. "Clinimetrics vs. psychometrics: an unnecessary distinction." *Journal of Clincial Epidemiology* 56: 1142-45.

[30] Teresi, J. A., R. R. Golden, P. Cross, B. Gurland, M. Kleinman, and D. Wilder. 1995. "Item Bias in Cognitive Screening Measures: Comparisons of elderly white, afro-american, hispanic and high and low education subgroups." *Journal of Clinical Epidemiology* 48(4): 473-83.

[31] Tesio, L., C. V. Granger, and R. C. Fiedler. 1997. "A unidimensional pain/disability measure for low-back pain syndromes." *Pain* 69: 269-78.

[32] Tsutakawa, R. K. and H. Y. Lin. 1986. "Bayesian estimation of item response curves." *Psychometrika* 51: 251-67.

[33] van Alphen, A., R. Halfens, A. Hasman, and T. Imbos. 1994. "Likert or Rasch? Nothing is more applicable than good theory." *Journal of Advanced Nursing* 20: 196-201.

[34] Ware, J. E. J. P., J. B. M. D. P. Bjorner, and M. M. A. Kosinski. 2000. "Practical Implications of Item Response Theory and Computerized Adaptive Testing: A Brief Summary of Ongoing Studies of Widely Used Headache Impact Scales." *Medical Care* 38(9 Suppl. II): 73-82.