Published in final edited form as:
Spine (Phila Pa 1976). 2008 May 20; 33(12): 1378–1383. doi:10.1097/BRS.0b013e3181732acb.

Letting the CAT out of the Bag: Comparing Computer Adaptive Tests and an Eleven-Item Short Form of the Roland-Morris Disability Questionnaire

Karon F. Cook, PhD[1], **Seung W. Choi**[2,3], **Paul K. Crane, MD MPH**[4], **Richard A. Deyo, MD MPH**[4], **Kurt L. Johnson, PhD**[1], and **Dagmar Amtmann, PhD**[1]

[1] Department of Rehabilitation Medicine, University of Washington, Seattle, WA

[2] Center on Outcomes, Research and Education, Evanston Northwestern Healthcare, Evanston, IL

[3] Northwestern University Feinberg School of Medicine, Chicago, IL

[4] Department of Medicine, University of Washington, Seattle, WA

Abstract

Study Design—A post-hoc simulation of a computer adaptive administration of the items of a modified version of the Roland Morris Disability Questionnaire.

Objective—To evaluate the effectiveness of adaptive administration of back pain-related disability items compared to a fixed 11-item short form.

Summary of Background Data—Short form versions of the Roland Morris Disability Questionnaire have been developed. An alternative to paper-and -pencil short forms is to administer items adaptively so that items are presented based on a person's responses to previous items. Theoretically, this allows precise estimation of back pain disability with administration of only a few items.

Materials and Methods—Data were gathered from two previously conducted studies of persons with back pain. An item response theory model was used to calibrate scores based on all items, items of a paper-and-pencil short form, and several computer adaptive tests (CATs).

Results—Correlations between each CAT condition and scores based on a 23-item version of the Roland Morris Disability Questionnaire ranged from 0.93 to 0.98. Compared to an 11-item short form, an 11-item CAT produced scores that were significantly more highly correlated with scores based on the 23-item scale. CATs with even fewer items also produced scores that were highly correlated with scores based on all items. For example, scores from a five-item CAT had a correlation of 0.93 with full scale scores. Seven- and nine-item CATs correlated at 0.95 and 0.97, respectively. A CAT with a standard-error-based stopping rule produced scores that correlated at 0.95 with full scale scores.

Conclusions—A CAT-based back pain-related disability measure may be a valuable tool for use in clinical and research contexts. Use of CAT for other common measures in back pain research, such as other functional scales or measures of psychological distress, may offer similar advantages.

Keywords

Health Status; Item response theory; Measurement; Outcomes; Quality of Life

Key points

- Because self-reports of back pain disability are used in clinical research studies where response burden is a critical issue, outcome instruments are needed that maximize measurement precision while minimizing response burden.
- Computer adaptive testing (CAT) was simulated using the items of the modified Roland and Morris Disability Questionnaire (RM-MOD).
- CATs with as few as five items produced scores that were highly correlated with item response theory (IRT) calibrated scores of the RM-MOD (RM-MOD$_{IRT}$). Correlations for various CAT conditions ranged from 0.93 to 0.98.
- The correlation between an 11-item CAT and RM-MOD$_{IRT}$ scores was 0.98. This correlation was significantly higher ($p<0.001$) than the correlation of 0.93 between scores from the 11-item fixed form and the RM-MOD$_{IRT}$.

INTRODUCTION

Self-reports of back pain-related disability often are used as endpoints in clinical research studies that evaluate interventions and treatments. Such measures also may supply information useful in clinical decision making. Among the most widely used measures of back pain disability is the Roland-Morris Disability Questionnaire (RM).[1] The RM is comprised of 24 items to which respondents answer “yes” or “no” to indicate limitations they experience due to back pain. The original RM was developed by identifying items of the Sickness Impact Profile (SIP)[2] relevant to low back pain. Patrick and colleagues[3] developed a 23-item, modified version of the RM (RM-MOD) by selecting items sensitive to change over time. The two versions share 19 items in common, but five of the original items were replaced by four new items in the modified version. Evidence regarding the psychometric properties of the RM-MOD has been published.[3]

The RM-MOD’s 23 items constitute a relatively brief scale and if it were the only instrument administered to patients it would not impose a substantial response burden. However in practice back pain-related disability is often one of a large number of patient-reported outcomes of interest, and typically several outcome instruments are administered.[4] Using traditional psychometric methods, comprehensive assessment would require completion of long questionnaires. This is suboptimal since response rates have been found to be significantly lower with longer versus shorter surveys.[5]

At least three short forms of the RM have been developed. The items included in the RM, the RM-MOD, and the three short forms are displayed in Table 1. Stratford and colleagues[6] developed an 18-item version (RM-18) based on classical psychometric analyses including calculation and evaluation of response frequencies, inter-item and item-total correlations, and coefficient alpha. They concluded that the RM-18 was as effective as the full 24-item RM. Atlas and colleagues constructed a 12-item questionnaire (RM-12).[7] The RM-12 was found to have somewhat lower coefficient alpha values (in the derivation sample, 0.82 for the RM-12 compared to 0.90 for the RM) but had reproducibility equal to that of the RM. The RM-12 also demonstrated high construct validity. Stroud and colleagues[8] developed an 11-item version of the RM (RM-11) using an item response theory approach. The coefficient alpha of the RM-11 was similar to that of the 24-item (0.88 versus 0.90, respectively). Correlations between RM-11 scores and those obtained with the RM-18 and 24-item measures were 0.95 and 0.93, respectively. The results of the three studies suggest that the RM may be shortened substantially without greatly sacrificing the strength of its psychometric properties.

Another approach to decreasing the number of items to which subjects respond is computer adaptive testing (CAT) in which, after a person responds to a starting item, items are selected and presented based on preliminary estimates of persons' trait level. This preliminary estimate is based on persons' responses to previous items.[9] After responses to each successive item, the preliminary estimate is updated and a new item is selected based on this estimate. Thus, the assessment is *adapted* to respondents' levels of the outcome being measured. Items that contribute substantively to estimating persons' trait levels are presented. As an example, if a patient responds to a question and indicates inability to walk a block, there is little point in asking about ability to walk a mile. Though CAT has been employed in educational and psychological testing for 30 years, only recently has it been applied to the measurement of health outcomes.[10–17] The purpose of the current study was to evaluate whether a CAT administration of back pain disability items could result in more efficient measurement of back pain disability.

METHODS

Data

Data were gathered from two previously conducted studies.[18, 19] IRB permission was obtained from the University of Washington Human Subjects Committee for this reanalysis of the data. Detailed methods have been published.[18, 19] One of the two studies was a multi-center prospective cohort study of 495 patients with presumed discogenic back pain ("the Discogenic study"[18]). Participants had one- or two-level disc degeneration confirmed by imaging and neurological evaluations. The second study was a clinical trial of 380 participants with low back pain randomly assigned to rapid magnetic resonance imaging or standard radiographs (the Seattle Lumbar Imaging Project, "SLIP").[19] In both studies, RM scores served as the primary outcome, and in both, data were collected at more than one time point. For the current study, we used only data collected at baseline.

Analyses

IRT Assumptions—The mathematical model that allows CAT is Item Response Theory (IRT). IRT is a probability model and estimation of scores is achieved without the requirement that every person respond to the same items.[17, 20] IRT models assume that a single latent construct (e.g., "back-pain related disability") drives persons' item responses (unidimensionality assumption). Health outcomes are conceptually complex and never perfectly meet the unidimensionality assumption.[21–24] The pertinent question is whether the presence of secondary dimensions cause the results of a unidimensional IRT calibration to be invalid.[25]

To evaluate the degree to which the unidimensionality assumption was met, we conducted a first-order, confirmatory factor analysis using Mplus software.[26] Because of the categorical nature of the response data, the polychoric correlation matrix was analyzed. In addition we plotted eigenvalues (scree plot) and evaluated the correlation between factors in a two-factor solution in an exploratory factor analysis.

Calibration to an Item Response Theory Model—We modeled RM-MOD item responses to the two-parameter logistic model (2-PL) using Parscale version 4.1.[27] The 2-PL model estimates both an item difficulty and an item discrimination parameter. Expected a posteriori score estimation was used so that scores could be estimated for persons who endorsed all (or none) of the items. Fit to the 2-PL model was calibrated using the computer macro, IRTFIT.[28] We report S-X^2 and S-G^2 fit statistics ($p<0.01$).[29, 30]

CAT Study Conditions and Simulations—With traditional measures, assessment stops when a respondent completes all items. With CAT measures, participants respond only to a subset of items, and so a “stopping rule” must be specified. Stopping rules can be based either on number of items (fixed-length CAT) or a standard error of measurement (SEM) can be specified (variable CAT). The SEM is an estimate of a measurement’s precision. It estimates the standard deviation of the differences between persons’ “true scores” and their observed scores (the ones obtained on the measure). With variable CATs, the administration continues until the pre-specified SEM is reached or all items have been administered.

We simulated 5-, 7-, 9-, and 11-item fixed-length CATs. We also simulated a CAT based on a SEM stopping rule of 0.5. All respondents in the current study answered all 23 items of the RM-MOD (RM-MOD$_{IRT}$). We used a computer algorithm to simulate CAT administration of the items. The computer program was written using SAS/STAT software, Version 9.1 for Windows XP-Pro.[31] In the simulation Item 1, “I stay at home most of the time because of my back” was “presented” to each respondent as the initial item, and an initial estimate was made of persons’ levels of back pain disability based on their responses. The item chosen to be presented next was the remaining item that provided the most “information” given the initial estimate of the person’s trait level (note: information in IRT is an extension of the concept of reliability.[32]) Based on response to this item, the estimate of a person’s trait level was updated. This process continued until the stopping rule was reached or all 23 items had been presented.

For purpose of comparison, in addition to the CAT simulations we calibrated IRT scores for the RM-11 (RM-11$_{IRT}$). These scores were compared to IRT scores based on the full RM-MOD (RM-MOD$_{IRT}$) and to scores obtained in the CAT conditions.

Evaluations of Study Conditions—The success of the CAT administrations and of the IRT-calibration of the RM-11 was evaluated by comparing them to the full-scale score (RM-MOD$_{IRT}$). Pearson product moment correlations were calculated between sets of scores. In addition, we calculated residuals. We defined residuals as CAT score (or RM-11 score) minus RM-MOD$_{IRT}$. These residuals indicated how far off the mark CAT and short form scores were from the 23-item RM-MOD$_{IRT}$ scores, our gold standard for the current study.

Though a scale may be more precise at some levels of back pain-related disability than others (e.g, moderate disability versus severe disability), with classical methods a single summary reliability estimate is calculated for the entire scale.[20] Thus differences in a scale’s measurement precision at different levels of trait are masked. An advantage of IRT models is that reliability is calculated for every level of trait. To have a reference point with which to compare the magnitude of residuals in the current study, we compared each person’s residual with the SEM obtained for that person in the 2-PL calibration of the 23-item scale. We then compared, across condition, the percentage of scores whose residuals were less than one SEM. The choice of one SEM was somewhat arbitrary.

RESULTS

Sample Characteristics

In the combined samples of the Discogenic study[18] and the SLIP,[19] 740 participants (85%) were white, 71 (9%) African-American or Black, 18 (2%) Asian, and 24 (3%) were of Hispanic origin. The mean age was 47 years (range = 18 to 93 years, SD = 13 years). Of the participants in the combined samples, 45.0% reported working full-time; 10.3% worked part-time; 20.5% were retired; and 4.2% were unemployed. Eighteen percent reported receiving workers’ compensation because of their back. Additional demographic and clinical details have been published elsewhere.[18, 19, 33]

Tests of IRT Assumptions

The results of a confirmatory factor analysis of a one-factor model (to evaluate unidimensionality) were assessed by examining several fit indices including the comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean residuals (SRMR). Strict standards for these measures have been suggested (CFI>0.95, RMSEA <0.06, SRMR <0.08).[25] In practice such high standards are seldom met in the context of health outcome measurement.[21, 34] Such was the case in the current study. The CFI was 0.90; RMSEA was 0.08; and the SRMR was 0.09.

When fit statistics fail to meet strict standards for good fit, Reeve and colleagues[25] recommend conducting an exploratory factor analysis, plotting eigenvalues against the rank of the eigenvalue (scree plot), and evaluating several statistics including percent of score variability on the first factor (≥20% is desirable), ratio of first and second eigenvalues (>4 supports unidimensionality assumption), and correlations among factors. Following this advice, we conducted a follow-up exploratory factor analysis. By these more relaxed standards, there was good support for the essential unidimensionality of the item responses. The scree test supported a single dominant dimension; the first factor accounted for 51.3% of the variance; the ratio of the first and second eigenvalues was 7.1; and the correlation between the first two factors was 0.69.

Fit to the 2-PL Model

Values of S-X^2 and S-G^2 statistics indicated good fit to the 2-PL model. None of the 23 items were found to be misfitting base on a criterion of p<0.01. The fit of the items to the model corroborated our judgment that the responses were essentially unidimensional and therefore appropriate for calibration using an IRT model.

Correlations among Scores

CATs with Fixed Numbers of Items—As expected, scores from CAT administrations with more items were more highly correlated with RM-MOD_{IRT} scores than were scores from CAT administrations with fewer items. Correlations between CATs with 5, 7, 9, and 11 items were 0.93, 0.95, 0.97, and 0.98, respectively. The correlation between RM-11_{IRT} scores and RM-MOD_{IRT} scores was 0.93. Coincidentally, this is exactly the value of the Pearson product-moment correlation reported by Stroud and colleagues for the correlation between summed RM-11 scores and summed scores of the 24-item version of the RM.[8] In comparison, the correlation between the 11-item fixed-length CAT scores and RM-MOD_{IRT} scores was 0.98. This correlation was significantly higher than the correlation of 0.93 obtained between scores from the 11-item fixed form and the RM-MOD_{IRT} (t=21.796, p<0.001). [35]

SEM-Based CAT—SEM-based CAT scores had a correlation of 0.95 with RM-MOD_{IRT} scores. The number of items administered in the SEM CAT condition ranged from 2–23 (See Figure 1). Though a mean of 8.3 items was administered, there was substantial variation in the number of items, and the majority of persons received seven or less items. For 121 of the 874 respondents (14%), all 23 available items were administered without the estimate reaching the stopping rule of SEM ≤ 0.50. This is indicative of the sparseness of items in the bank that target substantial portions of the sample. Indeed the RM-MOD exhibited substantial ceiling effects in the study sample; 96 persons (11%) scored in the upper 5% of the score range, endorsing 22 or 23 out of the 23 items. In contrast, for persons well-targeted by the scale, considerable efficiency was achieved as evidenced by the fact that in 53% of cases, five or fewer items were administered before the stopping rule of SEM ≤ 0.50 was reached.

Comparison of Residuals and SEMs

Figure 2 plots, for each study condition, the percentage of scores whose absolute residual (CAT/RM-11 score – RM-MOD$_{IRT}$ score) was less than one SEM (obtained for each score in the IRT calibration). Higher percentages indicate CATs that more successfully approximated the score obtained on the full scale. Percentages ranged from 68% for the 5-item CAT to 92% for the 11-item CAT. Notable is the substantially larger percentage of 11-item CAT scores (92%) compared to the RM-11 (73% within one SEM). In fact, better results were obtained with a 7-item CAT (76% of scores within one SEM) than with the RM-11, even with the 36% “savings” in response burden. A 7-item scale represents a 70% savings in response burden over the 23-item RM-MOD.

The SEM-based CAT had a stopping rule of SEM = 0.5, roughly equivalent to reliability of 0.69. This CAT condition performed better than the 5-item CAT, but not as well as the 7-item CAT.

DISCUSSION

The results of the current study indicate that substantial savings in response burden could be obtained using an adaptive approach to scaling back pain-related disability. CAT administrations with as few as 5 items predicted RM-MOD scores with reasonable accuracy, and an 11-item CAT performed substantially better than an 11-item short form. These results are evidence that the hypothesized scaling efficiency of computer adaptive testing can be realized in a clinical outcomes context. Of particular note is that these results were achieved with an item bank that was not developed specifically for computer adaptive administration. To be maximally effective, CAT item banks should be large and of sufficient breadth so that floor and ceiling effects are avoided.[9] In the current study, 11% of participants endorsed either all (6.5%) or all but one (4.5%) of the RM-MOD items, suggesting an insufficient number of items that target high levels of back pain-related disability—a ceiling effect. An item bank developed specifically for CAT not only would include more than 23 items but also more items designed to discriminate among persons with high levels of back pain-related disability. A CAT supported by such an item bank would be expected to result in even greater measurement efficiency across a broader range of disability than was observed in the current study.

The complexity of developing a CAT is acknowledged. It requires substantial effort in item development as well as advanced psychometric skills and specialized software. However, once a CAT is developed, the algorithm that selects items and estimates scores operates “behind the scene.” For responders the experience of reporting outcomes using CAT is no different from responding to a more traditional computer administered questionnaire, except that the number of questions and length of time required are reduced.

Even in contexts in which response burden is not a major issue, CAT retains advantages over assessment using fixed forms. As was suggested by our comparison of the 11-item CAT and the RM-11, with the same number of items, greater measurement precision can be achieved. A legitimate goal in its own right since greater measurement precision also results in smaller sample size requirements for clinical research since increases in precision result in increases in statistical power.[36]

A limitation of this study is that the CAT administrations were simulated. Participants did not respond to items of an actual CAT. We made the assumption that persons would give the same answers to the items of the RM-MOD whether they were presented by CAT or in a fixed, paper-and-pencil format. Though this assumption seems reasonable, its validity should be tested in future research. We also did not evaluate any of the non-technical, but critical issues regarding use of CAT in patient populations including whether patients are comfortable using a computer reporting their outcomes.

Our findings suggest that a CAT-based back pain-related disability measure could be a valuable tool for use in clinical and research contexts, particularly when response burden is a concern and/or multiple assessments are planned.

Acknowledgments

Dr. Deyo's effort was supported by a NIAMS Multidisciplinary Clinical Research Center, grant No. 5 P60-AR48093

Drs. Cook, Crane, Johnson, and Amtmann were funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant 5U01AR052171-03 to University of Washington. Information on the "Dynamic Assessment of Patient-Reported Chronic Disease Outcomes" can be found at http://nihroadmap.nih.gov/clinicalresearch/index.asp

References

1. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine 1983;8(2):141–4. [PubMed: 6222486]
2. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19(8):787. [PubMed: 7278416]
3. Patrick DL, Deyo RA, Atlas SJ, et al. Assessing health-related quality of life in patients with sciatica. Spine 1995;20(17):1899–908. [PubMed: 8560339]
4. Mannion AF, Elfering A, Staerkle R, et al. Outcome assessment in low back pain: how low can you go? Eur Spine J 2005;14(10):1014–26. [PubMed: 15937673]
5. Edwards P, Roberts I, Clarke M, et al. Methods to increase response rates to postal questionnaires. Cochrane Database Syst Rev 2007;(2):MR000008. [PubMed: 17443629]
6. Stratford PW, Binkley JM. Measurement properties of the RM-18. A modified version of the Roland-Morris Disability Scale. Spine 1997;22(20):2416–21. [PubMed: 9355224]
7. Atlas SJ, Deyo RA, van den Ancker M, et al. The Maine-Seattle back questionnaire: a 12-item disability questionnaire for evaluating patients with lumbar sciatica or stenosis: results of a derivation and validation cohort analysis. Spine 2003;28(16):1869–76. [PubMed: 12923478]
8. Stroud MW, McKnight PE, Jensen MP. Assessment of self-reported physical activity in patients with chronic pain: development of an abbreviated Roland-Morris disability scale. J Pain 2004;5(5):257–63. [PubMed: 15219257]
9. Cook KF, O'Malley KJ, Roddey TS. Dynamic assessment of health outcomes: time to let the CAT out of the bag? Health Serv Res 2005;40(5 Pt 2):1694–711. [PubMed: 16179003]
10. Andres PL, Black-Schaffer RM, Ni P, Haley SM. Computer adaptive testing: a strategy for monitoring stroke rehabilitation across settings. Top Stroke Rehabil 2004;11(2):33–9. [PubMed: 15118965]
11. Cook KF, Roddey TS, O'Malley KJ, Gartsman GM. Development of a Flexilevel Scale for use with computer-adaptive testing for assessing shoulder function. J Shoulder Elbow Surg 2005;14(1 Suppl S):90S–94S. [PubMed: 15726093]
12. Dijkers MP. A computer adaptive testing simulation applied to the FIM instrument motor component. Arch Phys Med Rehabil 2003;84(3):384–93. [PubMed: 12638107]
13. Forker JE, McDonald ME. Methodologic trends in the healthcare professions: computer adaptive and computer simulation testing. Nurse Educ 1996;21(4):13–4.
14. Haley SM, Ni P, Hambleton RK, et al. Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. J Clin Epidemiol 2006;59 (11):1174–82. [PubMed: 17027428]
15. Haley SM, Ni P, Ludlow LH, Fragala-Pinkham MA. Measurement precision and efficiency of multidimensional computer adaptive testing of physical functioning using the pediatric evaluation of disability inventory. Arch Phys Med Rehabil 2006;87(9):1223–9. [PubMed: 16935059]

16. Haley SM, Raczek AE, Coster WJ, et al. Assessing mobility in children using a computer adaptive testing version of the pediatric evaluation of disability inventory. Arch Phys Med Rehabil 2005;86 (5):932–9. [PubMed: 15895339]

17. Wiener A, Marcus E, Mizrahi J. Objective measurement of knee extension force based on computer adaptive testing. J Electromyogr Kinesiol 2007;17(1):41–8. [PubMed: 16497516]

18. Deyo RA, Mirza SK, Heagerty PJ, et al. A prospective cohort study of surgical treatment for back pain with degenerated discs; study protocol. BMC Musculoskelet Disord 2005;6:24. [PubMed: 15913458]

19. Jarvik JG, Hollingworth W, Martin B, et al. Rapid magnetic resonance imaging vs radiographs for patients with low back pain: a randomized controlled trial. JAMA 2003;289(21):2810–8. [PubMed: 12783911]

20. Embretson, SE.; Reise, SP. Item Response Theory for Psychologists. Mahway, NJ: Lawrence Erlbaum Associates, Publishers; 2000.

21. Cook KF, Teal CR, Bjorner JB, et al. IRT health outcomes data analysis project: an overview and summary. Qual Life Res. 2007

22. Reise SP, Haviland MG. Item response theory and the measurement of clinical change. J Pers Assess 2005;84(3):228–38. [PubMed: 15907159]

23. Reise SP, Waller NG, Comrey AL. Factor analysis and scale revision. Psychol Assess 2000;12(3): 287–97. [PubMed: 11021152]

24. Lai JS, Crane PK, Cella D. Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. Qual Life Res 2006;15(7):1179–90. [PubMed: 17001438]

25. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45(5 Suppl 1):S22–31. [PubMed: 17443115]

26. Muthen, BO.; Muthen, LK. Mplus User's Guide. Los Angeles, CA: Muthen & Muthen; 2001.

27. Muraki, E.; Bock, RD. PARSCALE 3: IRT based test scoring and item analysis for graded items and rating scales. Chicago, IL: Scientific Software International, Inc.; 1997.

28. Bjorner, JB.; Smith, KJ.; Orlando, M., et al. IRTFIT: A Macro for Item Fit and Local Dependence Tests under IRT Models. Lincoln, RI: QualityMetric Incorporated; 2006.

29. Orlando M, Thissen D. Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. Applied Psychological Measurement 2000;24(1):50–64.

30. Orlando M, Thissen D. Further investigation of the performance of S - X2: An item fit index for use with dichotomous item response theory models. Applied Psychological Measurement 2003;27:289–298.

31. The SAS System for Windows [computer program], Version 9.1. Cary, NC: SAS Institute; 2003.

32. Hambleton, R.; Swaminathan, H.; Rogers, HJ. Fundamentals of item response theory. Newbury Park, CA: Sage Publishing, Inc.; 1991.

33. Crane PK, Cetin K, Cook KF, et al. Differential item functioning impact in a modified version of the Roland-Morris Disability Questionnaire. Qual Life Res. 2007

34. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. Psychological Assessment 1999;7:286.

35. Steiger J. Ttests for comparing elements of a correlation matrix. Psychological Bulletin 1980;87(2): 245–251.

36. Viswanathan, V. Measurement Error and Research Design. University of Illinois, Urbana-Champaign; Sage Publications, Inc: 2005.

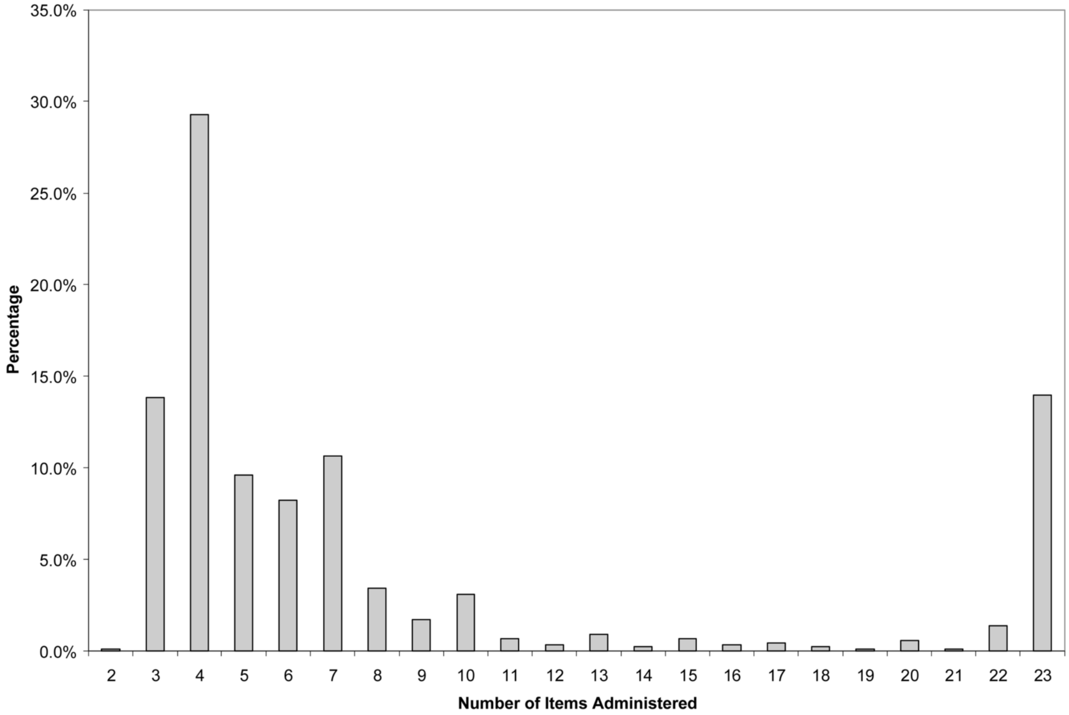

Figure 1.

Percentage of Cases in the SEM CAT Condition Receiving Different Numbers of Items.

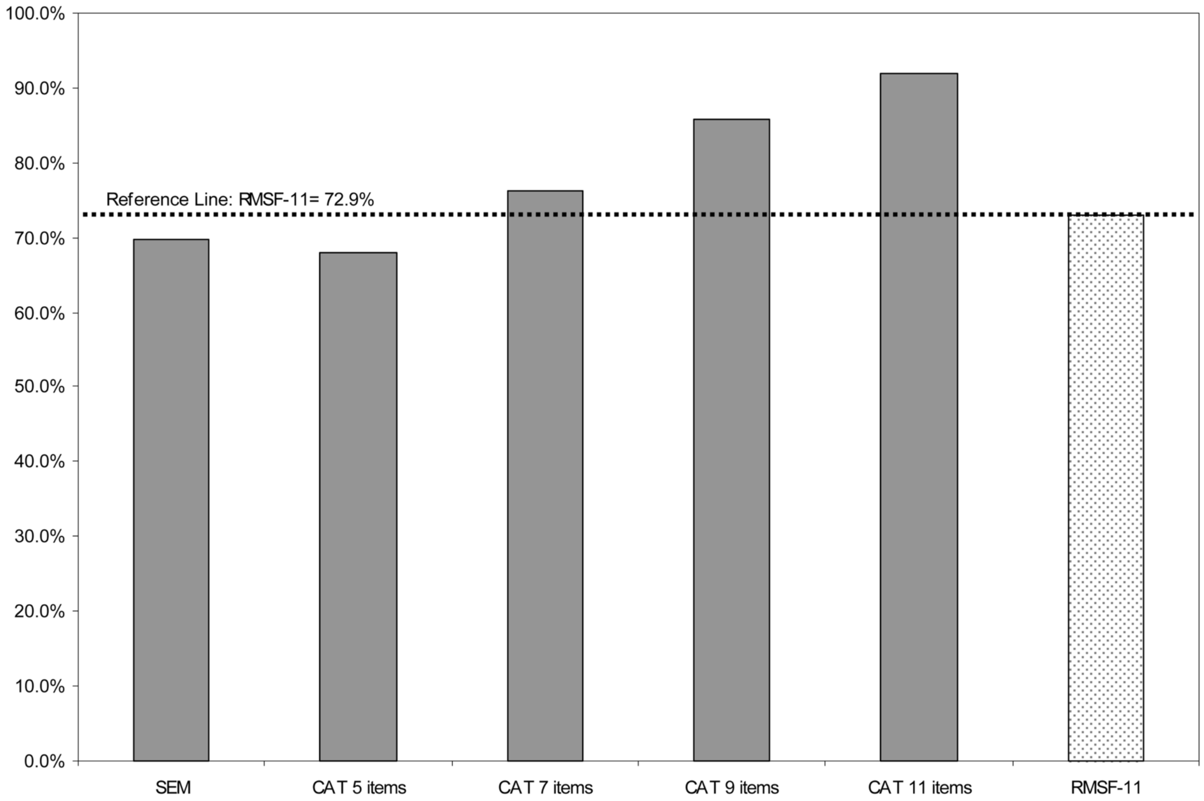

Figure 2.
Comparison of Percentage of Residuals within Study Condition that Are within One Standard Error of Measurement.

Table 1

Items of Five Versions of the Roland Morris Back Disability Questionnaire

Item	Original Version (RM)	Modified Version (RM-MOD)	18-item Version (RM-18)	12-item Version (RM-12)	11-item Version (RM-11)
I stay at home most of the time because of my back.	X	X	X		
I change positions frequently to try to get my back more comfortable.	X	X		X*	
I walk more slowly than usual because of my back.	X	X	X		X
Because of my back, I am not doing any of the jobs that I usually do around the house.	X	X	X		
Because of my back, I use a handrail to get upstairs.	X	X	X	X	X
Because of my back, I lie down to rest more often	X		X		
Because of my back, I have to hold onto something to get out of an easy chair	X	X	X		X
Because of my back, I try to get other people to do things for me	X		X		
I get dressed more slowly than usual because of my back.	X	X	X	X*	X
I only stand for short periods of time because of my back.	X	X	X	X*	X
Because of my back, I try not to bend or kneel down.	X	X	X	X	X
I find it difficult to get out of a chair because of my back.	X	X	X	X*	X
My back is painful almost all of the time	X	X	X	X*	
I find it difficult to turn over in bed because of my back.	X	X	X		
My appetite is not very good because of my back.	X				
I have trouble putting on my socks (or stockings) because of the pain in my back.	X	X	X		X
I only walk short distances because of my back pain.	X	X			X
I sleep less well because of my back.	X	X	X	X	
Because of my back pain, I get dressed with the help of someone else.	X				
I sit down for most of the day because of my back.	X				
I avoid heavy jobs around the house because of my back.	X	X	X		X
Because of back pain, I am more irritable and bad tempered with people than usual.	X	X	X		
Because of my back, I go upstairs more slowly than usual.	X	X	X		X
I stay in bed most of the time because of my back.	X	X		X*	

Item	Original Version (RM)	Modified Version (RM-MOD)	18-item Version (RM-18)	12-item Version (RM-12)	11-item Version (RM-11)
Because of my back problem, my sexual activity is decreased		X		X	
I keep rubbing or holding areas of my body that hurt or are uncomfortable		X		X	
Because of my back, I am doing less of the daily work around the house than I would usually do		X		X	
I often express concern to other people about what might be happening to my health		X			

* A reference to leg pain was added to these items.