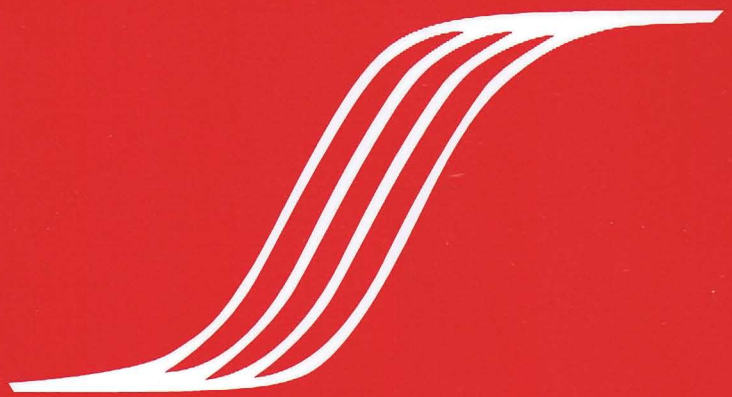
Volume 5, Number 1, 2001/2002 ISSN 1090-655X

Journal of Outcome Measurement®

Dedicated to Health, Education, and Social Science



REHABILITATION FOUNDATION INC.
Est. 1993
Research & Education

Vol.5, No. 1, 798-864 JOURNAL OF OUTCOME MEASUREMENT® 2001/2002

This issue of the
Journal of Outcome Measurement
was generously donated by
Alan Tennant

EDITOR

Richard F. Harvey. M.D. . . . . Rehabilitation Foundation, Inc.

ASSOCIATE EDITORS

Benjamin D. Wright . . . . . University of Chicago
Carl V. Granger . . . . . State University of Buffalo (SUNY)

HEALTH SCIENCES EDITORIAL BOARD

David Cella . . . . . Evanston Northwestern Healthcare
William Fisher, Jr. . . . . Louisiana State University Medical Center
Anne Fisher . . . . . Colorado State University
Gunnar Grimby . . . . . University of Goteborg
Perry N. Halkitis . . . . . New York University
Mark Johnston . . . . . Kessler Institute for Rehabilitation
David McArthur . . . . . UCLA School of Public Health
Tom Rudy. . . . . University of Pittsburgh
Mary Segal . . . . . Moss Rehabilitation
Alan Tennant . . . . . University of Leeds
Luigi Tesio . . . . . Foundazione Salvatore Maugeri, Pavia
Craig Velozo . . . . . University of Florida

EDUCATIONAL/PSYCHOLOGICAL EDITORIAL BOARD

David Andrich . . . . . .Murdoch University
Trevor Bond . . . . . .James Cook University
Ayres D'Costa . . . . . Ohio State University
George Engelhard, Jr. . . . . Emory University
Robert Hess . . . . . Arizona State University West
J. Michael Linacre . . . . . .MESA Press
Laura Knight-Lynn. . . . . .Rehabilitation Foundation, Inc.
Geofferey Masters . . . . . Australian Council on Educational Research
Carol Myford . . . . . Educational Testing Service
Nambury Raju . . . . . Illinois Institute of Technology
Randall E. Schumacker . . . . . .University of North Texas
Mark Wilson . . . . . .University of California, Berkeley

JOURNAL OF OUTCOME MEASUREMENT®

Volume 5, Number 1 2001/2002

Reviewer Acknowledgement

Articles

Comparison of Seven Different Scales used to Quantify Severity of Cervical Spondylotic Myelopathy and Post-Operative Improvement 798
A Singh, HA Crockard

The Impact of Rater Effects on Weighted Composite Scores UnderNested and Spiraled Scoring Designs, Using the Multifaceted Rasch Model..... 819
Husein M. Taherbhai and Michael James Young

The following article from Volume 4, Issue #3 is being reprinted due to errors in printing the tables:

Measuring Disability: Application of the Rasch Model to Activities of Daily Living (ADL/IADL)..... 839
T. Joseph Sheehan, Laurie M. DeChello, Ramon Garcia, Judith Fifield, Naomi Rothfield, Susan Reisine

Call for Papers.....864

REVIEWER ACKNOWLEDGEMENT

The Editor would like to thank the members of the Editorial Board who provided manuscript reviews for the Journal of Outcome Measurement, Volume 5, Number 1.

JOURNAL OF OUTCOME MEASUREMENT®, 5(1), 798-818
Copyright© 2001, Rehabilitation Foundation, Inc

Comparison of Seven Different Scales used to Quantify Severity of Cervical Spondylotic Myelopathy and Post-Operative Improvement

A Singh

HA Crockard

Department of Surgical Neurology

National Hospital for Neurology and Neurosurgery, London UK

Considerable uncertainty exists over the benefit that patients receive from surgical decompressive treatment for cervical spondylotic myelopathy (CSM). Such difficulties might be addressed by accurate quantification of CSM severity as part of a trial determining the outcome of surgery in different patient groups. This study compares the applicability of various existing quantitative severity scales to measurement of CSM severity and the effects on severity of surgical decompression. Scores on the following scales were determined on 100 patients with CSM preoperatively and then again six months following surgical decompression: Odom's Criteria, Nurick grade, Ranawat grade, Myelopathy Disability Index (MDI), Japanese Orthopaedic Association (JOA) Score, European Myelopathy Score (EMS) and Short Form-36 Health Survey (SF36). All the scales showed significant improvement following surgery. However, each had differing qualities of reliability, validity and responsiveness that made them more or less suitable. The MDI showed the greatest sensitivity between different severity levels, sensitivity to operative change and reliability. However, analysis of all the questionnaire scales into components that looked at different aspects of function revealed potential problems with redundancy and a lack of consistency. This prospective observational study provides a rational basis for determining the advantages and disadvantages of different existing scales in measurement of CSM severity and for making adaptations to develop a scale more specifically suited to a comprehensive surgical trial.

Requests for reprints should be sent to Alan Crockard, DSc., Department of Surgical Neurology, National Hospital for Neurology and Neurosurgery, Queen Square, London, WC1N 3BBG, UK

INTRODUCTION

Rational observation of disease management requires a consideration of and measurement of the outcome of such management. In this context, outcome may be defined as an "attributable effect of intervention or its lack on a previous health state." (Calman 1994). Information about the outcome of different treatments is important not only to clinicians, and to patients and their families, but in the current era of cost constraints, also to the health provider and the health purchaser. In the present climate of evidence-based health care, all clinicians in their individual practices must aspire to achieve comparable best results; such aims can only be realised by a proper consideration and quantification of the outcomes of their treatments.

Treatment of CSM well illustrates this increasing need for a more rigorous investigation of management outcomes. Decompressive surgery for cervical spondylotic myelopathy (CSM) was first performed by Victor Horsley in 1892 and has been a standard practice for many years. However, the selection of appropriate patients for such procedures and the determination of the correct stage in the disease to operate remains uncertain. In fact, Rowland (Rowland, 1992) has questioned the fact that surgery has any role in cervical spondylotic myelopathy, arguing that there has been no large prospective surgical series and that retrospective series in the literature (Phillips, 1973; Clarke and Robinson, 1956) do not demonstrate any treatment advantage over conservative management. While the lack of such data does not invalidate operative treatment, different clinicians do appear to vary greatly in their selection practices for decompressive surgery and it is likely that a considerable number of patients are unnecessarily operated upon, while others are operated upon too late or not at all. As discussed, the increasing demand for scientific justification of clinical practice makes some form of large prospective comparison of the outcomes for operated versus non-operated patients extremely timely.

Currently, clinicians rely on specific symptoms, such as difficulty with gait or urinary difficulties, together with specific findings on clinical examination and radiological imaging, to identify the most

severe forms of cervical spondylosis and to decide when surgery is appropriate. It is clear that more quantitative severity and outcome measures would be required for a clinical trial, and such measures might also ultimately prove useful in clinical assessment of individual patients.

A variety of quantitative assessment scales now exist that have or could potentially be applied to the quantification of CSM severity and so facilitate proper study of the outcome of surgery. The goal of our study was therefore to explore prospectively the applicability of various impairment, disability and handicap scales to CSM patients pre- and post-operatively and, if no one scale is found to be ideal, to determine those applicability and statistical qualities of different scales that would be desirable in the development of an ideal scale.

METHODS

Subjects

We prospectively studied 100 patients with CSM, who were consecutively referred and accepted for decompressive surgery to the Neurosurgical Unit at National Hospital for Neurology and Neurosurgery. The median age of the patients was 58 years and there were 62 males and 38 females. All patients had the diagnosis corroborated by MRI and none had undergone previous neck surgery or had any other pathology that might have resulted in functional impairment. Ethical committee approval and informed consent from each patient was obtained under the guidelines of the Hospital Policy. The patients were under the care of six Neurosurgeons. The assessor was a Nurse Practitioner previously experienced in the use of such scales (Singh and Crockard, 1999) who had no input in surgical decision-making.

Of the 100 patients, 50 anterior cervical discectomies (Cloward's or Smith Robinson's) and 50 posterior decompressions (laminectomies n=16, laminoplasties n=34) were performed by 7 different neurosurgeons.

Study design and data analysis

Each patient was assessed by the same assessor. Scores for the following functional assessment scales were determined shortly before surgery and then again 6 months after surgery:

1. Myelopathy Disability Index (MDI): this is a disability scale applied to assessment of rheumatoid myelopathy and constituting a shortened form of the Health Assessment Questionnaire (HAQ), which in turn is adapted from the Activities of daily living (ADL) scale. Scores range from 0 (normal) to 30 (worst) (Casey et al., 1996)
2. Japanese Orthopaedic Association Score (JOA): a disability scale that attempts to look at various impairment categories such as disability related to upper motor neurone, radicular and sphincter deficits. Scores range from 0 (worst) to 17 (normal) (Hirabayashi et al., 1981).
3. European Myelopathy Score (EMS): a scale adapted from the JOA for Western use that also includes pain assessment. Scores range from 5 (worst) to 18 (normal) (Herdman, et al., 1994).
4. Nurick Score: a simple scale mainly focusing on walking disability, ranging from 1 (normal) to 5 (worst) (Nurick, 1972).
5. Ranawat: a simple impairment scale, ranging from 1 (normal) to 4 (3B) (worst) (Ranawat, 1979).
6. Odom's criteria: a simple score looking at overall surgical outcome, ranging from 1 (best outcome) to 4 (no change or worse) (Odom, et al., 1958).
7. The MOS 36-item short-form health survey (SF36): A complex health questionnaire measuring disability and handicap (% of normal 100%) (Ware and Sherbourne, 1992).

These different outcome measures were then analyzed with respect to their properties of internal consistency, sensitivity, validity and responsiveness. Data were analysed statistically using the SPSS package version 9.

Figure 1

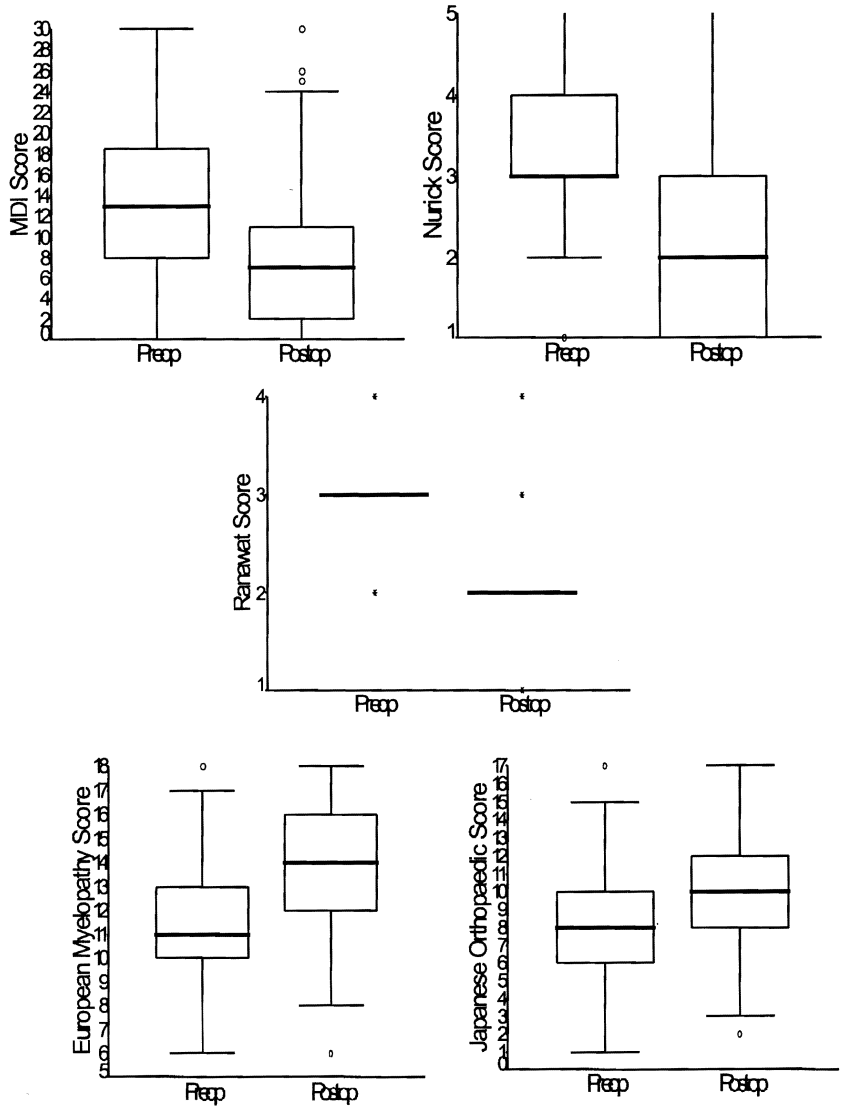


Figure 1: Box plots of the 100 pre-operative and 99 post-operative scores of all the patients on 5 different scales. (One patient died shortly following surgery.) For the MDI, the Nurick and the Ranawat scales a better score is a lower value, while for the EMS and JOA better scores are represented by higher values. The circles represent outlying values greater than 1 ½ interquartile intervals and the stars represent extremes greater than 3 interquartile intervals. In all cases, the improvement following surgery was statistically significant (Wilcoxon) (table 1).

Figure 2

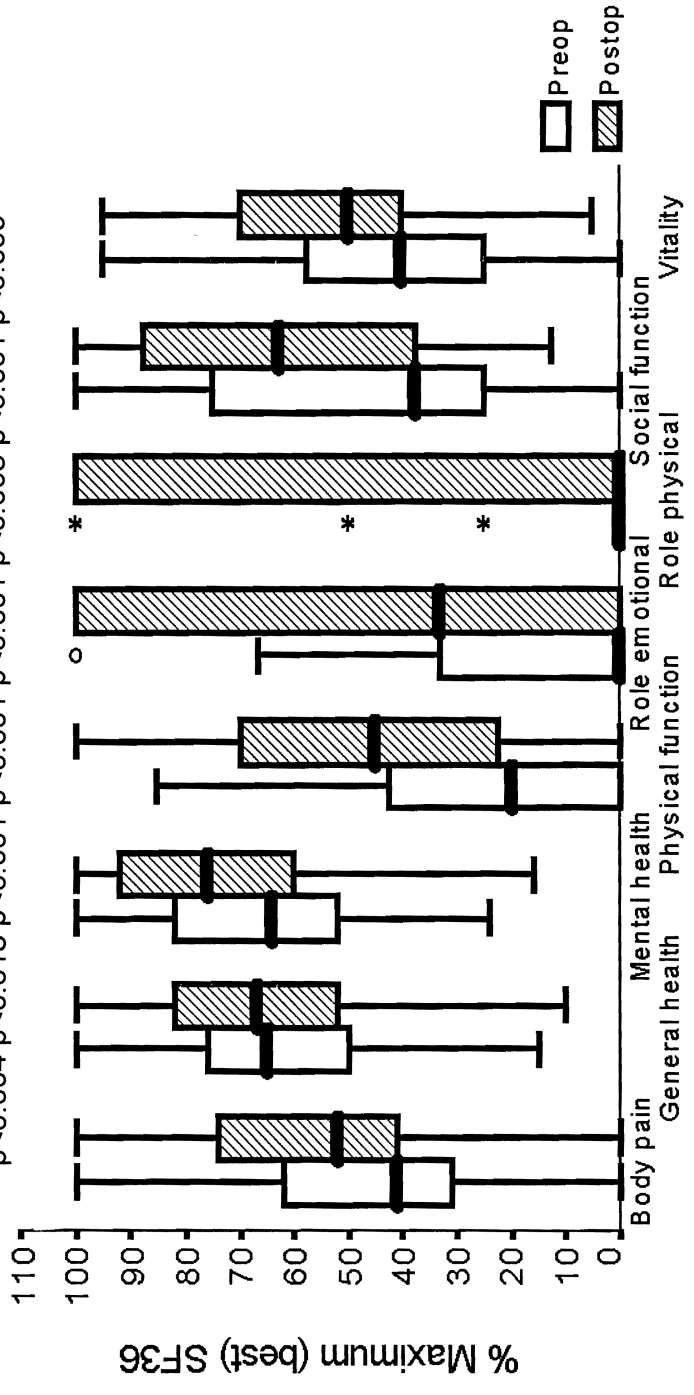


Figure 2: Box plots of pre and post operative scores for the 8 categories of the SF-36 Questionnaire. These scores have all been transformed to percentages for comparison, where 100 % is the best possible score. Each category shows significant improvement following surgery (Wilcoxon).

RESULTS

Patient and Operative Details

The median length of hospital stay for the 100 patients was 8 days and there was a 3 % wound infection rate. There was one peri-operative death due to cardio-respiratory failure 3 weeks following surgery. Thus only 99 comparisons were available.

Pre- and Post-operative Scale Scores

All scales recorded an improvement following surgery (Figures 1, 2). On a Wilcoxin test, this improvement was significant in each case (Table 1 and Figure 2 for SF 36 subcategories). Note that Odom's criteria only record operative results so there are no pre- and post-operative values. There were a minority of patients who scored worse 6 months following surgery (e.g. 8 out of 99 for the MDI). On each scale, these were slightly different patients (see correlations section).

Sensitivity to change

While all of the scales showed a statistically significant improvement following surgery, this does not reveal the magnitude of the change. It is clearly desirable for a scale to show a large sensitivity to change. This was quantified by calculating the Normalised Change, the mean of the differences following surgery for the 99 subjects (in whom a comparison was possible) divided by the overall median of the 199 pre- and post-operative scores, i.e. (mean of (preop score - postop score))/ median of all scores. The mean rather than median of differences was used because, while the scale values were not normally distributed, the *differences* in values did follow an approximately normal distribution. The MDI was found to be the best scale according to this criterion, while the EMS was the worst (Table 1).

Absolute Sensitivity

It may be desirable to have a high sensitivity to distinguish different

Table 1 **Comparison of properties of different scales. The significance of improvement is the p-value of the operative change. Sensitivity to change is mean of (preop score - postop score)/median of all scores. Coefficients of variation pre-op and post-op and the reliability (Cronbach's *a*) pre-op and post-op are also shown for all scales.**

SCALE	SIGNIFICANCE OF IMPROVEMENT	SENSITIVITY TO CHANGE	CO-EFFICIENT OF VARIATION PRE-OP	CO-EFFICIENT OF VARIATION POST-OP	INTERNAL CONSISTENCY (CRONBACH'S α PRE-OP)	INTERNAL CONSISTENCY (CRONBACH'S α POST-OP)
MDI	P<0.001	0.52	0.85	1.29	0.92 (11 SCORES) 0.76 (4 CATEGORIES)	0.95 (11 SCORES) 0.81 (4 CATEGORIES)
EMS	P<0.001	0.18	0.27	0.29	0.68	0.77
JOA	P<0.001	0.21	0.5	0.4	0.72 (6 SCORES) 0.66 (4 CATEGORIES)	0.73 (6 SCORES) 0.65 (4 CATEGORIES)
NURICK	P<0.001	0.42	0.33	1	-	-
RANAWAT	P<0.001	0.34	0	0	-	-
SF36™	P<0.001	0.32	0.41	0.68	0.82 (36 ITEMS)	0.86 (36 ITMES)

absolute levels of severity between patients in the sample group, as well as sensitivity to changes following surgery. Absolute sensitivity was quantified by the coefficient of variation (the interquartile range divided by the median). It is seen that the Ranawat score has poor sensitivity for distinguishing patients with different levels of severity, because the range across the patients is narrow. This is illustrated by the fact that the box plot shows a single horizontal line instead of a box (Figure 1). Thus nearly all pre-operative patients were scored at one level and post-operatively at a level one grade better, indicating that the Ranawat score nevertheless records a post-operative improvement.

The Nurick scale was found to have much greater sensitivity post operatively, perhaps indicating that the scale was more sensitive at distinguishing milder levels of severity.

Internal Consistency

If different questions in a multipart questionnaire are attempting to measure the same parameter e.g. CSM severity, then there should be consistent scoring within patients. This is measured by Cronbach's alpha (Cronbach and Meehl, 1955), a normalised measure of correlations between multiple components of a scale. A score of 1 indicates a perfect correlation. The very high Cronbach's alpha values of the MDI (table 1) show that the questionnaires were reliably completed, but also suggest the possibility of redundancy. When the 11 questions of the MDI were split into 4 categories (walking, hand function, transfers and dressing), the alpha scores were somewhat lower. This is appropriate, since if different questions within a questionnaire are designed to address different parameters, then it is not desirable to have high internal consistency.

Correlations of Scores

To explore the validity of the different scales, correlation coefficients were calculated for the pre-operative scores (Table 2A), post-operative scores (Table 2B) and for the changes following surgery (Table 2C). All correlations were corrected for the fact that some scales

recorded no disability as the maximum value, while others recorded no disability as the minimum value.

It was found that some scales were correlated better than others; the best correlation was found post-operatively between the MDI and the EMS scales (r= 0.82) which are both disability questionnaires, while the poorest correlation was postoperatively between the SF36 (measuring handicap and disability) and the Ranawat (measuring neurological impairment).

The correlations were poorer when comparing operative changes. Many values were close to zero or even negative.

Breaking down Scales into Components

The generally poor correlation between scales, with better correlation between more similar scales (e.g. the postoperative MDI and EMS scores) could be due to some scales measuring different aspects of function or impairment. This was initially investigated by empirically dividing the multi-part scales into components measuring certain aspects of disability or impairment. This breakdown might also reveal that different individual aspects have different potentials for improvement following surgery. Thus, the Normalised Changes, measuring the magnitude of operative change (sensitivity to change) of the different components of the three multipart disability questionnaires were calculated and compared (Table 3).

A reasonably consistent trend was apparent across the scales revealing that good improvement tended to occur in hand function as assessed by all three scales addressing this aspect, while both scales looking at sphincter function showed that it remained little changed by surgery. Within the SF 36, physical and social function and social role changed most (Figure 2), but no corroboration was available for these parameters since they were not measured by any other scale. The findings in general support the possibility that the poor correlations might be better if one compared specific aspects of CSM rather than overall scales. However, since the scale components have not been validated when looked at individually, one has to interpret differences in improvement between these specific aspects with cau-

Table 2A

	MDI	EMS	RANAWAT	NURICK	JOA	SF36
Pre-op MDI	1	-	-	-	-	-
Pre-op EMS	0.75	1	-	-	-	-
Pre-op RANAWAT	0.51	0.61	1	-	-	-
Pre-op NURICK	0.66	0.69	0.71	-	-	-
Pre-op JOA	0.56	0.62	0.47	0.59	1	-
Pre-op SF36™	0.48	0.42	0.31	0.38	0.40	1

Table 2B

	Post-Op MDI	Post-Op EMS	Post-Op RANAWAT	Post-Op NURICK	Post-Op JOA	Post-Op SF36
Post-Op MDI	1	-	-	-	-	-
Post-Op EMS	0.82	1	-	-	-	-
Post-Op RANAWAT	0.67	0.63	1	-	-	-
Post-Op NURICK	0.71	0.74	0.75	1	-	-
Post-Op JOA	0.57	0.72	0.42	0.51	1	-
Post-Op SF36™	0.35	0.35	0.25	0.36	0.37	1

Table 2C

	MDI Change	EMS Change	RANAWAT Change	NURICK Change	JOA Change	SF36 Change	ODOM 'S Change
MDI Change	1	-	- -	-	-	-	-
EMS Change	0.27	1	- -	-	-	-	-
RANAWAT Change	0.22	0.23	1 -	-	-	-	-
NURICK Change	0.32	0.32	0.55	1	-	-	-
JOA Change	0.15	0.35	0.02	0.19	1	-	-
SF36™ Change	0.22	0.12	0.003	0.13	0.28	1	-
'ODOM'S Change	0.02	0.27	0.33	0.25	0.24	0.19	1

Table 2A,B,C. Correlations of score pre-operatively (2A) post-operatively (2B) and operative changes i.e. differences between pre-operative and post-operative scores (2C).

tion. For example, the greater improvement in hand function after surgery might simply reflect a greater sensitivity of the questionnaires to this component rather than a genuinely greater improvement.

Correlations of Components

In order to seek some validation of the component sensitivities and to explore why the overall scale correlations of operative change were low, the next step was to perform correlations between these components in a similar way to the correlations performed above for the overall scales. Thus, the components of the multi-part scales questioning walking function were directly correlated with each other as well as with the Ranawat and Nurick scales (which have a one-dimensional measure primarily based on walking), while hand and blad-

Table 3 **Breakdown of scales into components sensitivity to change in these aspects fllowing surgery.**

	MDI sensitivity to change	EMS sensitivity to change	JOA sensitivity to change
WALKING	0.58	0.2	0.21
HAND	0.70	0.22	0.35
DRESSING	0.35	0.2	-
SPINCTER	-	0.03	0.04
WASHING/ TRANSFERS	0.42	-	-
PAIN	-	0.22	-
SENSORY LOSS	-	-	0.33

Table 3. Three scales were broken down into their component aspects and sensitivities to change recalculated for these separate components. For example, the JOA has questions relating to walking, hand and spincter function and sensory change. The hand function components recorded by these scales change much more than bladder-related components.

der components were similarly correlated between those scales that had aspects pertaining to these components (Table 4A, B, C).

It was found that, particularly for hand and bladder function improvement, correlations were still very poor. The correlation of operative changes for two apparently similar questions on the JOA and EMS, namely bladder function, was only 0.23. On analysing individual patient's responses, the inconsistencies were clear. For example, patient number 10 indicated his bladder became worse post-operatively on the EMS, going from normal to inadequate, but on the JOA he reported only a mild disturbance both pre- and post-operatively.

Table 4A

	MDI Walk Change	EMS Walk Change	RANAWAT Change	NURICK Change	JOA Walk Change
MDI Walk Change	1	-	-	-	-
EMS Walk Change	0.07	1	-	-	-
RANAWAT Change	0.26	0.25	1	-	-
NURICK Change	0.34	0.23	0.55	1	-
JOA Walk Change	0.13	0.48	0.19	0.29	1

Table 4B

	JOA Hand Change	MDI Hand Change	EMS Hand Change
JOA Hand Change	1	-	-
MDI Hand Change	0.12	1	-
EMS Hand Change	0.25	0.26	1

Table 4C

	EMS Bladder Difference	JOA Bladder Difference
EMS Bladder Difference	1	-
JOA Bladder Difference	0.23	1

Table 4A, B, C. Components such as walking, hand function, bladder were similarly correlated between those scales that had aspects pertaining to these components.

DISCUSSION

All the quantitative measures of CSM severity satisfied the most basic requirement of a scale useful in assessing the effects of surgery in that they were all able to demonstrate a significant improvement in score following surgery. This consistent finding is of course also indicative of a genuine benefit resulting from such intervention. However, such an effect would only be properly demonstrated by a study that included a period of follow up longer than 6 months and that included a comparison with a similar group of CSM patients that were not operated upon.

Sensitivities of Different Scales

While all the scales showed significant improvement following surgery, they have other properties that make them more or less suitable form assessment of CSM. The MDI is sensitive to change and also gives a wide range of absolute values, which means there is good sensitivity to differences between patients. On the other hand, the Ranawat score, while being sensitive to change, was very poor at distinguishing different levels of absolute severity. This study, in looking at both pre- and post-operative scores, thus illustrates the important point that it is insufficient to attempt validation of scales only on absolute measurements; their properties may be considerably different if the scales are also to be used to assess the effect of operative or other interventions. In addition, widely differing absolute sensitivities between pre- and post-operative measurements suggests that different scales may have different applicability to different patient groups. For example, the Nurick score had a much greater sensitivity post-operatively, suggesting a greater ability to distinguish between different levels of severity at the milder end of the scale.

Internal Consistency of Different Scales

The multi-part questionnaires had good internal consistency (internal reliability), particularly the MDI, suggesting that the question-

naires were being reliably completed. However, the high level of reliability may entail some redundancy where very similar questions concentrating on the same aspect of disability are asked repeatedly. Even worse, if different aspects of dysfunction are considered sometimes to be affected to different degrees in different patients, it would seem inappropriate that questions testing these different aspects always score too similarly. The lower alpha score when the MDI is divided into categories comparing different aspects does suggest some genuine effect in distinguishing these categories. Nevertheless, the presence of multiple questions within the same category, while not resulting in poorer sensitivity and sensitivity to change, does point to redundancy and therefore inefficiency. A glance at the questions of the MDI (**appendix**) reveals that it tends to ask repeated questions on a few limited categories of disability. After the initial demonstration of high internal consistency during an initial study, indicating that the patients answer the questions reliably, perhaps redundant questions could simply be removed when designing an ideal scale used in assessing CSM severity.

Intra-rater and inter-rater reliability were not investigated in this study. Since the MDI, EDM, JOA and SF 36 are patient rated, inter-rater reliability is irrelevant for such scales. Instead, internal consistency is a measure of reliability across questions within the questionnaire. The Ranawat and Nurick scores are simple and one-dimensional and have previously been shown to have good intra- and inter-rater reliability.

Correlations between Scales

Possible flaws in the scales are suggested when looking at correlations between the scores on the various scales. The concept of internal consistency does not necessarily imply validity and accuracy, i.e. whether or not a scale is actually measuring what it purports to measure (Wassertheil-Smoller, 1995). Scales are ideally validated by comparing them with a gold standard. This is most relevant when they are used as a convenient surrogate for a gold standard definitive investigation that is invasive, risky or cumbersome, or perhaps when

used to predict an outcome that eventually becomes clear over time. In the absence of a gold standard in CSM, the scales were simply correlated with each other to see if certain inconsistencies became apparent.

It was found that, while correlations between similar scales were sometimes high, correlations between recorded operative *changes* were poor. This is because change is likely to be a much more sensitive indicator of dissimilarities between scales. For example, if a patient generally scores well on different scales pre-operatively and there is only a small post-operative improvement, the changes may well be in different directions on the different scales, while the post-operative absolute scores all still remain generally high. These highlighted differences between scales could reflect aspects of change that some scales measure which others ignore. Thus, a mildly affected patient may generally score quite highly, but operative decompression might change certain aspects much more than others. This point again illustrates the importance of validating scales by looking at changes rather than confining assessment to patients in the static state.

Breakdown of Scales into Components

To explore the possibility that different scales measure different aspects of function, the individual scales were subdivided on empirical grounds into different functional components. There were indeed differences between components, with hand function showing the greatest improvement, walking showing moderate improvement and bladder function showing minimal improvement. However, these results must be interpreted with caution since they could reflect that different scales are simply better at measuring changes in different aspects of function rather than there being real differences in change of function. Indeed, when one actually correlates these different aspects of function by correlation of the components between the scales, the coefficients are often no better than for the overall scales, throwing doubt upon the validity of making strong inferences about the separate components of a scale. This finding also suggests that the poor

overall correlations of improvement between the overall scales cannot be explained on the basis that the different scales record different aspects of this improvement, but instead lead one to question the validity of some or all of the scales. On review of individual patients' responses, it is clear that apparently similar single questions are sometimes answered very differently in different scales, possibly due to the phrasing of such questions.

An important precept of a multi-part scale is that there is an overall unidimensionality i.e. "overall severity". Thus the scale simply adds all the different components from which patients with myelopathy might suffer. No hierarchy of components is considered at all, other than perhaps more questions being asked on areas that are more important for patient functioning. This study has addressed the relationship between the components of different scales and found that, particularly when looking at *changes* in severity, this unidimensionality cannot be applied - some components deteriorate while others improve and there is no consideration of which are more important.

CONCLUSIONS

An ideal scale should be as quantitative as possible and show good sensitivity between patients and sensitivity to change. It should also be scored reliably and be simple to use. Of the scales investigated, the MDI best reflects these characteristics. This scale constitutes a questionnaire that focuses upon a limited range of aspects of disability; the findings indicate that such a scale does not necessarily suffer in terms of sensitivity. Instead, repeated questioning on similar aspects of function may reflect redundancy. Moreover, the poor correlations between the operative changes recorded by the overall scales and their components indicates that repeated questions on different or even similar aspects of function may actually reveal considerable inconsistencies. Thus, while a scale such as the MDI appears to be adequate for a prospective outcome trial of intervention in CSM, it is possible that an ideal scale might be one that makes a simple single quantitative measurement on a limited aspect of function.

APPENDIX

MYELOPATHY DISABILITY INDEX

Please tick the response which best describes your usual abilities over the past week	Without **ANY** difficulty	With **SOME** difficulty	With **MUCH** Difficulty	**UNABLE** to do so
Score	**0**	**1**	**2**	**3**
Rising are you able to:				
Stand up from an armless straight chair?				
Get in and out of bed?				
Eating are you able to:				
Cut your meat?				
Lift a full cup or glass to your mouth?				
Walking are you able to:				
Walk outdoors on a flat ground?				
Climb up five steps?				
Hygiene are you able to:				
Wash and dry your entire body?				
Get on and off the toilet?				
Grip are you able to:				
Open jars which have been previously opened?				
Activities are you able to:				
Get in and out of the car?				
Dressing are you able to:				
Dress yourself, include tying shoelaces, and doing buttons on your shirt or blouse?				
TOTAL	A	B	C	D

Note: If aids or assistance from another is required to perform any of the tasks please score the activity as "with much difficulty". Total score = A+B+C+D (range 0-33). The final score is expressed as a percentage.

ACKNOWLEDGEMENTS

We would like to acknowledge Mr. Adrian Casey, Mr. William Harkness, Mr. Neil Kitchen, Mr. Michael Powell, Professor David Thomas and Mr. Lawrence Watkins for allowing us to study their patients.

REFERENCES

Calman, K.C. (1994). The ethics of allocation of scarce health care resources: a view from the centre. *J Med Ethics,* 1994 June; 20(2): 71-4.

Casey, A.T.H, Bland, J. M., and Crockard, H. A. (1996). 'Development of a functional scoring system for rheumatoid arthritis patients with cervical myelopathy', *Annals of the Rheumatic Diseases,* 55, 901-906.

Clarke, E., and Robinson, P.K. (1956). 'Cervical myelopathy: a complication of cervical spondylosis', *Brain*, 79, 483-510.

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Herdman, J., Linzbach, M., Krzan, M., et al. (1994). The European myelopathy score. In: Bauer BL, Brock M, Klinger M, eds. *Advances in Neurosurgery*. Berlin: Springer, 266-8.

Hirabayashi, K., et al. Operative results and post-operative progression of ossification among patients with cervical osteophytic posterior longitudinal ligament. *Spine* 1981; 6:354-364.

Nurick S. (1992). The Pathogenesis of the Spinal Cord Disorder As sociated with Cervical Spondylosis. *Brain*, **95**, 87-100.

Odom G.L., Finney W., and Woodhall B. (1958). Cervical disc lesions. JAMA; 166, 23 - 28.

Phillips, D.G. (1973). 'Surgical treatment of myelopathy with cervical spondy losis', *J Neurol Neurosurg Psychiatry,* 36, 879 - 884.

Ranawat, C., O'Leary, P., Pellici, P., et al. (1979). 'Cervical fusion in rheumatoid arthritis, *Journal of Bone and Joint Surgery America*, 61A, 1003-10.

Rowland L. P. (1992). Surgical treatment of cervical spondylotic myelopathy: time for a controlled trial. *Neurology*, 42(1): 5-13.

Singh, A., and Crockard, H. A. (1999) Quantitative assessment of cervical spondylotic myelopathy by a simple walking test. *Lancet.* 1999 Jul 31; 354(9176): 370-3.

Ware, J.E., and Sherbourne, C. D. (1992). The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. *Med Care*, 30:473-83.

Wassertheil-Smoller, S. (1995). *'Biostatistics and Epidemiology - A Primer for Health Professionals'*, Springer - Verlag New York, Inc.

JOURNAL OF OUTCOME MEASUREMENT®, 5(1), 819-838
Copyright© 2001, Rehabilitation Foundation, Inc

The Impact of Rater Effects on Weighted Composite Scores Under Nested and Spiraled Scoring Designs, Using the Multifaceted Rasch Model

Husein M. Taherbhai[1]
Learning Research and Development Center
University of Pittsburgh

Michael James Young
Harcourt Educational Measurement

Constructed-response or open-ended tasks are increasingly used in recent years. Since these tasks cannot be machine-scored, variability among raters cannot be completely eliminated and their effects, when they are not modeled, can cast doubts on the reliability of the results. Besides rater effects, the estimation of student ability can also be impacted by differentially weighted tasks/items that formulate composite scores. This simulation study compares student ability estimates with their true abilities under different rater scoring designs and differentially weighted composite scores. Results indicate that the spiraled rater scoring design without modeling rater effects works as well as the nested design in which rater tendencies are modeled. As expected, differentially weighted composite scores have a confounding effect on student ability estimates. This is particularly true when open-ended tasks are weighted much more than the multiple–choice items and when rater effects interact with weighted composite scores.

FOOTNOTE

[1]Authors' names appear in alphabetical order.

Requests for reprints should be sent to Husein M. Taherbhai, University of Pittsburgh, Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara St., Room 802, Pittsurgh, PA 15260.

INTRODUCTION

Constructed response or open-ended tasks have been increasingly used in assessments in recent years. Since these tasks cannot be machine-scored, trained raters are used to score them. In a crossed design where every rater scores every task of every examinee, the recovery of the examinee abilities in simulation studies is very accurate (Hombo, Thayer & Donoghue, 2000). However, because of time and cost considerations, it is impossible for every rater to rate all examinees. When fully crossed designs are not used, variability among raters cannot be completely eliminated, and when rater effects are not modeled, biased ability estimates can result. Hombo and her colleagues, however, found that spiraled designs for assigning raters to the scoring of tasks performed better than nested designs in reducing the bias of ability estimates.

Besides rater effects, the use of composite scores that apply a priori weights to items and tasks from different formats (e.g., the College Board's *Advanced Placement* examinations (College Board, 1988)) can also have an effect in the estimation of student ability. These composite scores can have a confounding effect on examinee abilities when they interact with rater effects (Taherbhai & Young, 2000).

Taherbhai and Young (2000) used data from the Reading: Basic Understanding section of the New Standards English Language Arts (ELA) Examination to study the interaction of rater effects with composite scores. The New Standards ELA Examination consisted of both multiple- choice items (MC) and open-ended (OE) tasks. The data were used to form different weighted composite scores, which were then analyzed for rater effects using the multifaceted Rasch model. Results indicated that the interaction of rater effects with the weighted composite scores could dramatically alter the estimates of student abilities.

This study examines the impact of rater effects on weighted composite scores under nested and spiraled scoring designs. Raters are modeled to reflect a nested design (i.e., raters scoring all tasks across a subset of examinees), and a spiraled design, (i.e., raters scor-

ing a subset of tasks across all examinees) across a complex structure of differentially weighted composite scores. Examinee ability is then modeled using the multifaceted Rasch model.

The primary purpose of this study was to examine how well the ability parameters of examinees are recovered under different rating designs using conditions of rater effects and differentially weighted composite scores. The study also examined the effects of raters on composite scores for student classifications based on cutpoints.

Design and Methodology

Various log linear models can be used to analyze the hypothesis of rater and weight effects in scoring the open-ended sections of an examination. One such model in Item Response Theory, is the multifaceted Rasch model which can provide information on examinees, items, raters, and their interactions for ordered response categories (Linacre, 1989). The resulting probabilistic equation for a modified rating scale model (Andrich, 1978) incorporating the different measurement facets (i.e., examinees, raters, and items) can be presented in logarithmic form as:

$$\log\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \beta_n - \delta_i - \lambda_j - \tau_k \qquad (1)$$

where

P_{nijk} = probability of examinee n being rated k on item i, by rater j,

P_{nijk-1} = probability of examinee n being rated k-1 on item i, by rater j,

β_n = ability of examinee n,

δ_i = difficulty of item i,

λ_j = severity of rater j,

τ_k = difficulty in rating step k relative to step k-1.

The parameters of this model can be estimated using the FACETS program of Linacre (1989).

PROCEDURE

Simulated Data Generation

Data incorporating rater effects under the many-faceted Rasch model were simulated. Response data were generated for a test consisting of 20 dichotomously scored multiple-choice (MC) items and 3 open-ended (OE) (i.e., constructed response tasks) each scored on a 0 to 4 rubric. The data set consisted of 12 equally spaced true student ability (or thetas) from –2.00 to 2.00. Each of these true thetas was used to create 1,000 sets of examinee responses, giving a total examinee sample size of 12,000. Besides true student abilities, the following parameters were included in the generation of data:

1. Item Difficulty Parameters: The twenty multiple-choice item difficulty parameters were selected so that they were evenly spaced in the interval from –2.00 to 2.00. The three open-ended tasks were simulated with item difficulty parameters of -1.00, 0.00, and 1.00 respectively. The four step parameters on a rubric of 0 to 4 were kept constant across the open-ended items at –1.00, -0.33, 0.33, and 1.00.

2. Rater Parameters: The three rater parameters that were used for this study were –0.50, 0.00, and 0.50. The size of these rater parameters reflects those seen in Taherbhai and Young (2000).

Scoring Designs

Two scoring designs, nested and spiraled, were considered in this study. In the nested condition, each rater scores all tasks for a subset of the examinees. Under this design, biased estimates of student ability can occur, depending on the combination of rater with student ability. For example, a lenient rater whose tendency is to award a higher score than what the examinee actually deserves could have a group of examinees of high ability. Similarly, a severe rater whose tendency is to award a lower score than what the examinee actually deserves could rate a group of examinees of low ability.

As Hombo et al. (2000) explain, extreme raters tend to pull

the Item Response Function (IRF) in the direction of the rater's rating tendency. That is, lenient raters tend to shift the IRF to the left or down the ability scale while severe raters tend to shift the IRF to the right or up the ability scale. By the same token, moderate raters (those who are neither too lenient nor too severe) should have a moderate effect on student ability estimates.

Under the spiraled design, each rater scores only a subset of tasks for all of the examinees. However, the tasks are rotated so that raters are crossed with respect to students and tasks but rate only some tasks for some examinees and other tasks for other examinees. Under this condition, too, there are various rater and examinee combinations that could result in biased estimations of examinee ability.

Since this paper does not profess to exhaust all possible rater design of interest, two designs were selected to illustrate the results of certain patterns of assignment of raters.

Nested Design

Under this design, the lenient rater rated the lowest ability examinees across all replications, the moderate rater rated the moderate ability examinees across all replications, and the most severe rater rated the highest ability examinees across all replications. Since each rater rated four students whose performances were replicated 1,000 times on the OE tasks, the total number of ratings performed by each rater was 12,000.

This combination of raters and examinees was selected because according to Hombo, et al.'s (2000) results, this nested design (their Nested Design 1) was the one those that showed the greatest deviation in the recovery of the true ability parameters for extreme scores when rater effects were ignored. Furthermore, in our case, this combination of raters and examinees would also give us a chance to examine the effects of moderate raters on examinee ability estimates.

Spiraled Design

Under this design, each examinee was rated by a different rater on

on each of the three tasks. Specifically, the first examinee's first task was scored by the lenient rater, their second task by the moderate rater, and their third task by the severe rater. For the second examinee, the assignment of raters to tasks rotated so that the moderate rater scored the first task. For the third examinee, the rater assignment rotated yet again, so that the most severe rater scored the first task. This rotation of scoring pattern was then repeated for the remaining students. These ratings were awarded, as in the nested de-

Table 1 **Assignment of Raters in Nested and Spiraled Designs**

Examinees	Nested Designs			Spiraled Designs		
	OE Task 1	OE Task 2	OE Task 3	OE Task 1	OE Task 2	OE Task 3
1	Rater 1	Rater 1	Rater 1	Rater 1	Rater 2	Rater 3
2	Rater 1	Rater 1	Rater 1	Rater 2	Rater 3	Rater 1
3	Rater 1	Rater 1	Rater 1	Rater 3	Rater 1	Rater 2
4	Rater 1	Rater 1	Rater 1	Rater 1	Rater 2	Rater 3
5	Rater 2	Rater 2	Rater 2	Rater 2	Rater 3	Rater 1
6	Rater 2	Rater 2	Rater 2	Rater 3	Rater 1	Rater 2
7	Rater 2	Rater 2	Rater 2	Rater 1	Rater 2	Rater 3
8	Rater 2	Rater 2	Rater 2	Rater 2	Rater 3	Rater 1
9	Rater 3	Rater 3	Rater 3	Rater 3	Rater 1	Rater 2
10	Rater 3	Rater 3	Rater 3	Rater 1	Rater 2	Rater 3
11	Rater 3	Rater 3	Rater 3	Rater 2	Rater 3	Rater 1
12	Rater 3	Rater 3	Rater 3	Rater 3	Rater 1	Rater 2

Note: Rater 1 is Lenient, Rater 2 is Moderate, Rater 3 is Severe.

sign, across all replication for each student performance. Under this design, too, each rater rated every student on 1,000 performance replications for a total of 12,000 ratings. This particular spiral design was a modification of Hombo, et al's., (2000) Spiral Design 1. The details of assigning raters to tasks and examinees for the nested and spiraled designs are shown in Table 1.

Weights Assigned to Items and Tasks

In order to analyze the impact of differentially weighting multiple-choice items and open-ended tasks, four different composite scores were created. Each composite was created with the method used by the Advanced Placement Program (College Board, 1988), where the part scores for multiple-choice items and open-ended tasks are assigned weights to create a composite score with a target set of percentages for a fixed total.

The first composite represented a baseline condition where equal weights were used for the part scores: that is, the parts are "naturally weighted" with respect to MC and OE totals. Composites 1, 2, and 3 used different part weights to produce scores with multiple-choice to open-ended contributions of 75%-to-25%, 50%-to-50%, and 25%-to-75% respectively in the total scores.

These different composite scores were used with rater effects for the nested and spiral designs discussed above, and also without the inclusion of rater effects, to produce various scoring designs.

Data Analysis

In order to establish a common metric for comparing the different composite scores, all of the item, task, and item-step difficulty parameters were considered fixed across different designs and weighting conditions. The calibrations of the composites were anchored to these item/task parameter estimates when producing Rasch parameter estimates pertaining to examinee ability, rater severity, and OE tasks' step difficulty. The Rasch ability estimates were found for each composite score point on each weighted composite scale under each of the two scoring designs discussed above.

The FACETS program uses Unconditional Maximum Likelihood estimation (UCON) for estimating examinee abilities. In this study, examinee ability parameters used to generate data for examinee responses were considered as true examinee abilities. The examinee abilities that became infinite due to a zero or perfect score were adjusted by a fractional score adjustment of 0.3: Ability estimates corresponding to zero scores were instead estimated for a score of 0.3, while abilities for perfect scores were instead estimated for the maximum obtainable raw score minus 0.3. With respect to the weights assigned to items and tasks, the FACETS program uses the assigned weights as mutiplicative factors of the scores earned by the examinees.

As in Hombo, et al. (2000), two measures of accuracy in estimating examinee abilities were examined, the squared bias and the mean squared error (MSE). These measures were calculated as follows:

$$Bias^2 = \left(\bar{\theta}_{estimated} - \theta_{true}\right)^2$$

$$MSE = \frac{\sum_{n=1}^{1000}\left(\theta_{estimated} - \theta_{true}\right)^2}{1000}$$

Cutpoints at the quartiles of the Rasch ability estimates for the various composite scores without rater effects were compared with each other and then with the modeling of rater effects to examine changes in student classifications across the different composites.

RESULTS

Table 2 presents true examinee ability parameters, the squared bias, and the mean squared errors, across the two designs and across the two rater effects conditions, i.e., modeled and not modeled.

Parameter Recovery for Baseline Weighting

Under the nested design without modeling rater effects, the expected pattern of examinee ability estimates was generally observed. That is, when rater effects were not modeled, the impact of the severe rater was to move the estimates of examinees at the higher ability levels downward and for the lenient rater to move the estimates of examinees at the lower abilities upward. The impact of ignoring rater tendencies under the nested design is clear from Figure 1, which show the estimated vs. true ability estimates for the modeled and non-modeled rater effects under the baseline weighting. When rater effects were modeled, the ability estimates were pulled much closer to a 45-degree line (see Figure 2).

Both the squared bias and the mean square errors are reduced for extreme raters when they are modeled in the nested design.

Under the spiraled design there is a slight overall increase in bias from the nonmodeled rater condition to the modeled rater condition. However, the difference at each ability level is fairly small so that the two ability estimate plots coincide through most of the ability distribution (see Figures 3 and 4). Overall the mean square errors are slightly higher under the spiraled design when raters are modeled.

When the spiraled design is compared to the nested design under the nonmodeled rater condition, both the squared bias and the mean square errors are decreased. When raters are modeled, there is an overall increase in the mean square errors from the nested to the spiraled design. However, there is no noticeable difference in the squared bias between the two designs.

In general, the squared bias under the spiraled design with nonmodeled raters is lower than for the nested design with nonmodeled raters, and is comparable to modeled rater conditions. These results concur with the finding of Hombo, et al. (2000) that support the notion that when ratings are spiraled, rater effects get spread across examinees and the recovery of ability estimates is fairly accurate. However, the spiraled design did show slightly larger mean square errors than the nested design when raters were modeled.

Table 2 **Estimation of Squared Bias, and Mean Square Effects**

True Theta	Rater Effects Not Modeled							
	Nested Design				Spiraled Design			
	Baseline	Comp. 1	Comp. 2	Comp. 3	Baseline	Comp. 1	Comp. 2	Comp. 3
-2.00								
Sqd. Bias	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00
MSE	0.38	0.40	0.41	0.67	0.40	0.42	0.43	0.70
-1.64								
Sqd. Bias	0.01	0.00	0.02	0.05	0.01	0.01	0.01	0.04
MSE	0.26	0.27	0.31	0.53	0.28	0.29	0.31	0.57
-1.27								
Sqd. Bias	0.02	0.01	0.04	0.09	0.01	0.01	0.01	0.03
MSE	0.22	0.22	0.26	0.42	0.22	0.23	0.24	0.43
-0.91								
Sqd. Bias	0.04	0.02	0.07	0.14	0.00	0.00	0.01	0.01
MSE	0.21	0.20	0.25	0.41	0.16	0.18	0.17	0.24
-0.55								
Sqd. Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MSE	0.17	0.18	0.18	0.29	0.16	0.17	0.17	0.27
-0.18								
Sqd. Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MSE	0.17	0.18	0.19	0.29	0.17	0.18	0.19	0.29
0.18								
Sqd. Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MSE	0.16	0.17	0.18	0.27	0.14	0.15	0.14	0.20
0.55								
Sqd. Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
MSE	0.17	0.19	0.18	0.28	0.16	0.18	0.18	0.27
0.91								
Sqd. Bias	0.04	0.02	0.06	0.11	0.00	0.00	0.00	0.01
MSE	0.22	0.21	0.26	0.42	0.19	0.20	0.20	0.34
1.27								
Sqd. Bias	0.03	0.01	0.04	0.09	0.01	0.00	0.01	0.01
MSE	0.22	0.23	0.26	0.42	0.18	0.21	0.19	0.30
1.64								
Sqd. Bias	0.01	0.00	0.02	0.05	0.01	0.01	0.02	0.06
MSE	0.27	0.29	0.31	0.51	0.31	0.33	0.34	0.62
2.00								
Sqd. Bias	0.01	0.00	0.02	0.03	0.01	0.01	0.01	0.06
MSE	0.36	0.38	0.40	0.63	0.40	0.43	0.42	0.75

Errors Across Designs for Nonmodeled and Modeled Rater

Rater Effects Modeled							
Nested Design				Spiraled Design			
Baseline	Comp. 1	Comp. 2	Comp. 3	Baseline	Comp. 1	Comp. 2	Comp. 3
0.01	0.01	0.01	0.03	0.02	0.02	0.02	0.07
0.37	0.40	0.40	0.65	0.42	0.45	0.47	0.83
0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.05
0.24	0.25	0.27	0.48	0.28	0.29	0.31	0.60
0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02
0.19	0.21	0.21	0.33	0.21	0.23	0.23	0.41
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.16	0.18	0.18	0.27	0.18	0.20	0.19	0.31
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.16	0.17	0.18	0.28	0.16	0.17	0.17	0.27
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.17	0.18	0.18	0.28	0.17	0.17	0.18	0.28
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.15	0.16	0.17	0.26	0.16	0.17	0.18	0.28
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.17	0.18	0.18	0.27	0.16	0.17	0.17	0.26
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
0.18	0.19	0.19	0.30	0.19	0.20	0.20	0.34
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
0.19	0.21	0.20	0.32	0.20	0.22	0.22	0.38
0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.06
0.25	0.27	0.27	0.45	0.31	0.33	0.33	0.60
0.00	0.00	0.00	0.02	0.01	0.00	0.01	0.06
0.34	0.37	0.36	0.60	0.41	0.44	0.43	0.77

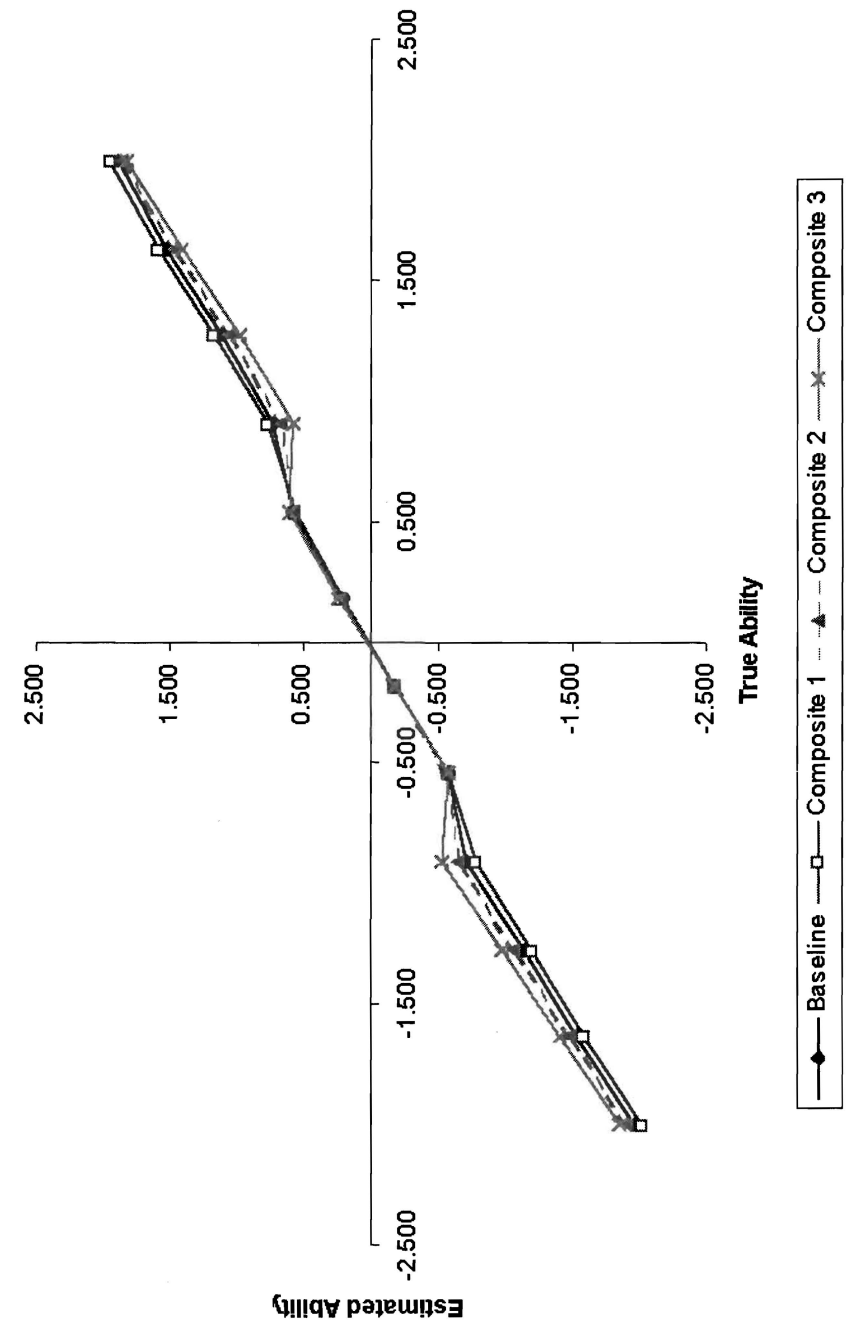


Figure 1: Estimated vs. True Ability: Weighted Nested Design Without Raters Modeled.

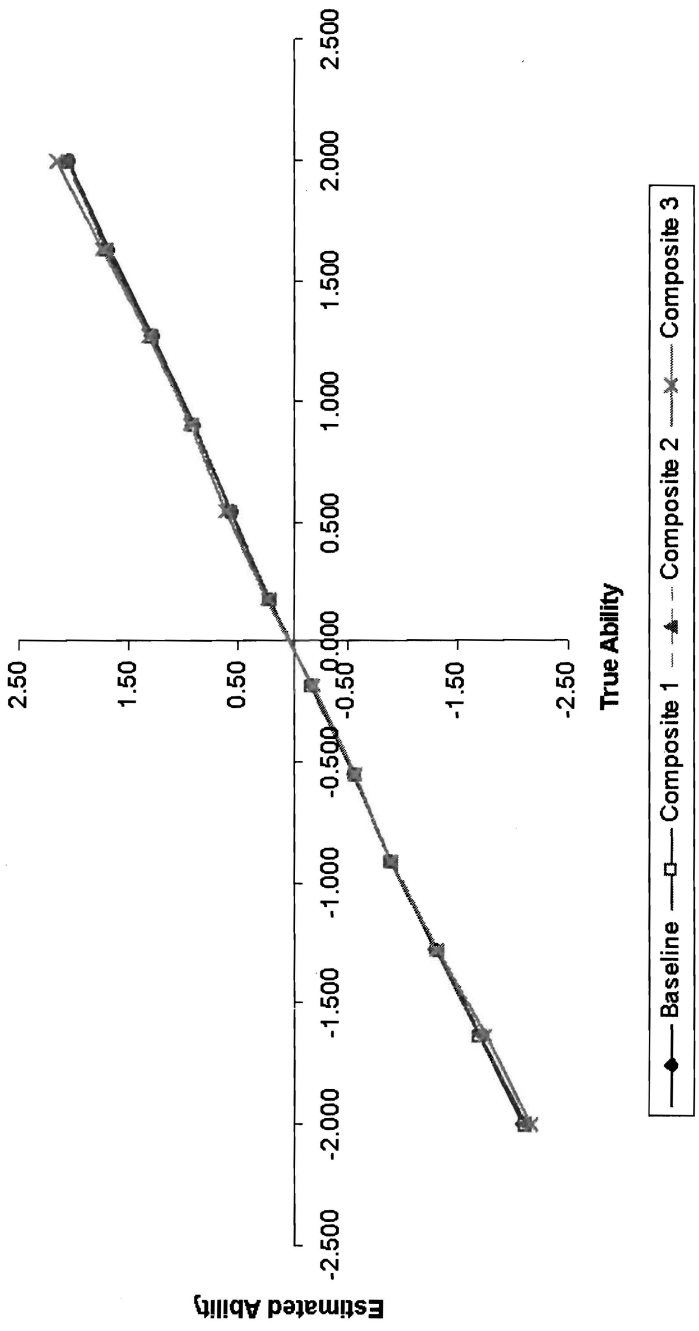


Figure 2: Estimated vs. True Ability: Weighted Nested Designs <u>With</u> Raters Modeled.

Parameter Recovery for Weighted Composites

Across the weighted composites, the mean squared errors increased as the weight assigned to the OE tasks increased. This was true for both the nested and spiraled designs, regardless of whether or not rater effects were modeled. The squared bias, on the other hand, was fairly constant except for the ends of the ability distribution in Composite 3.

For the nested design, when rater effects were not modeled and weights were included in the analysis, the rater effects interacted with the weights assigned to the composite. The composites with the highest OE task weights showed the greatest impact on the estimates of examinee ability at ends of the distribution (Figure 1). This is due to the effect of the extreme raters scoring examinees at the low and high ends of the ability distribution. When rater effects are modeled, the recovery of examinee abilities improves at the extreme ends of the ability distribution (Figure 2). Changes in OE task weights had little effect on the estimation of examinees of moderate abilities regardless of the modeling of raters, since these were the examinees scored by moderate raters.

Plots of the estimated vs. the true abilities for the spiraled design composites are shown for the non-modeled rater effects in Figure 3, and for the modeled rater effects in Figure 4. With the exception of Composite 3, the recovery of parameters was fairly accurate for the spiraled design regardless of whether the rater effects were modeled. The modeling of raters in Figure 4 had the effect of "straightening" the plots in Figure 3 except for Composite 3, which has the greatest weight placed on OE tasks of all of the composites. Contrary to the decrease in squared bias, the mean square errors increased slightly from the non-modeled rater condition to the modeled rater condition when the OE task weights in the composites increased.

Student Classification at the Quartiles for Different Composites

When cutpoints at the quartiles were examined (Table 3) under the two designs and the different weighting conditions, the results did

Figure 3: Estimated vs. True Ability: Weighted Spiraled Designs Without Raters Models

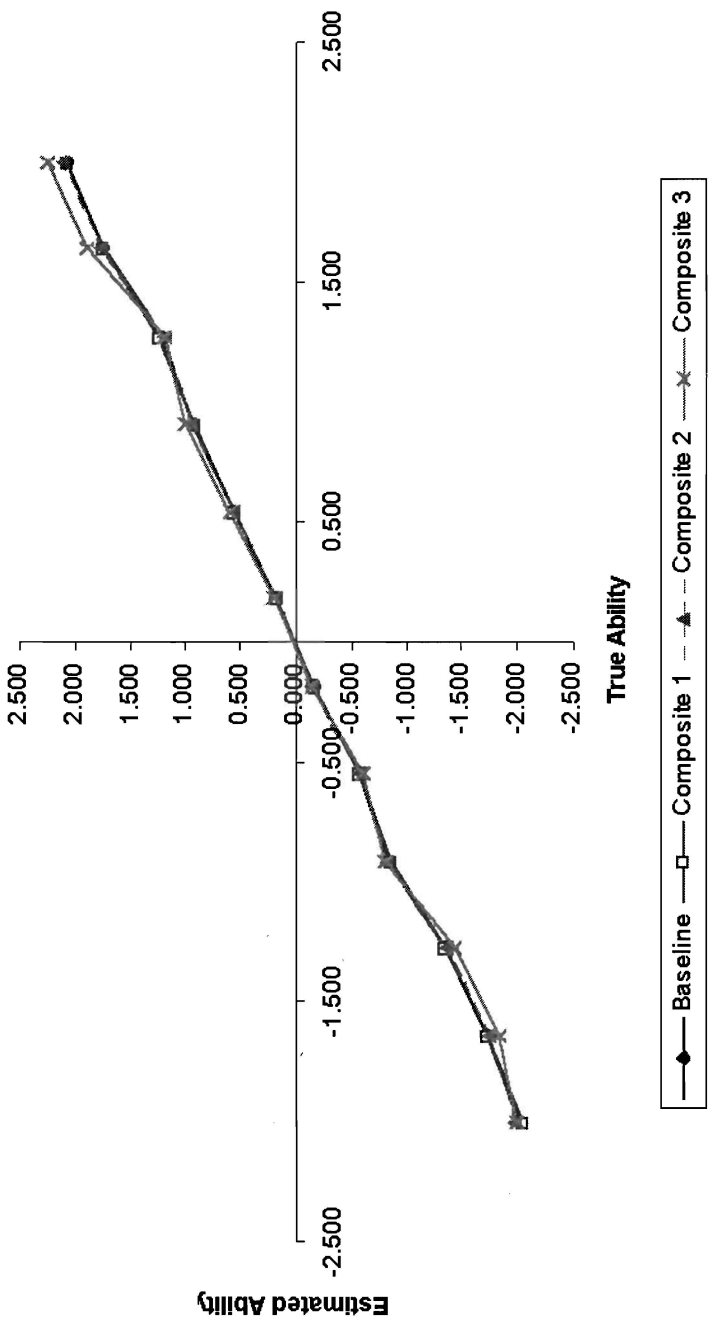
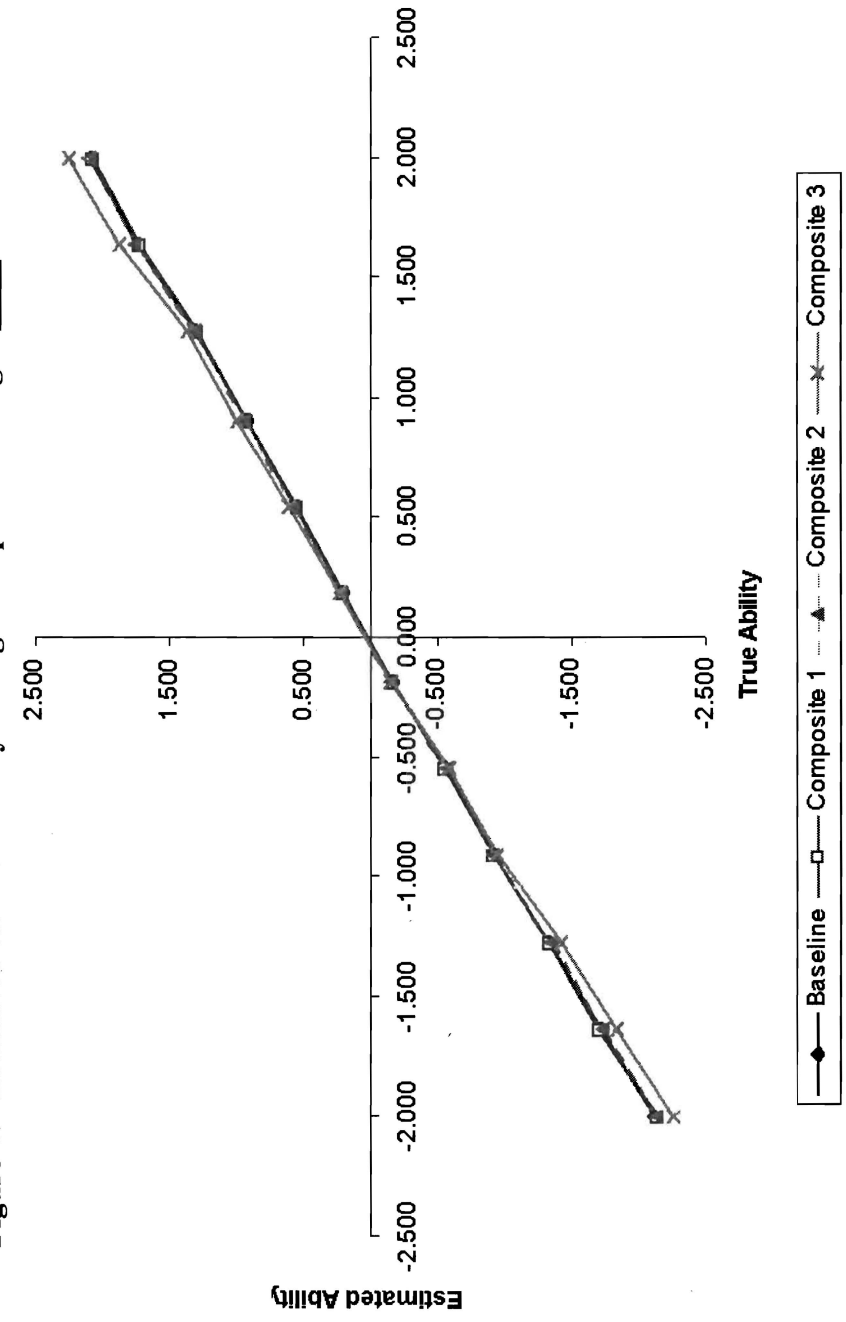


Figure 4: Estimated vs. True Ability: Weighted Spiraled Designs <u>With</u> Raters Models



not differ much from Taherbhai and Young's (2000) results. Here again, there was little movement from one condition to another under the two designs, and consistency in examinee classification was very high.

CONCLUSIONS

The impact of rater effects is well documented in the literature (Engelhard, 1994, 1996; Hombo, et al., 2000; Linacre, 1989; Lunz, et al., 1990). The composite effect of weighing MC items and OE tasks differentially, however, has not received much attention.

Under the nested design, when very severe or lenient raters are paired with examinees at the ends of the ability distribution, examinee ability estimates can be systematically distorted. As Lunz, Wright, and Linacre (1990) point out, if raters are used to score tasks for assessments and their effects are not modeled, it may lead to poor ability estimates especially if extreme raters are paired with examinees that are extreme in their abilities. Furthermore, weighing items and tasks differentially confounds rater effects and further complicates the recovery of true ability estimates of examinees.

The use of spiraled over nested designs when rater effects are not modeled, was justified in Hombo et al., (2000) due to the decrease in bias that these designs provided in estimating examinees' abilities. This result is supported by our study: The recovery of examinee abilities and the overall MSEs was better for spiraled designs than nested designs when rater effects where not modeled. The spiraled design with unmodeled rater effects was also comparable to the modeled nested design except for the very slight increase in MSE.

However, any advantage that the spiraled design has in reducing the bias of ability estimates decreased when composite scores were used that placed much greater weights on open-ended tasks than those on the multiple-choice items. In this situation, the bias and the MSE increased not only for the unmodeled nested design but also for both the modeled and the unmodeled spiraled designs.

As stated in the paper, this study is not exhaustive of all the possible rater designs and weighted conditions that could be included. Further research needs to be undertaken to examine the complex in-

Table 3 Percent of Students Changing Classification with Respect to Quartiles

	Nested Design		Spiralled Design	
	Classification Change		Classification Change	
Condition/Cutpoint	Up	own	Up	Down
Baseline				
Rater Effects Modeled vs. Not Modeled				
Q3	-	1	3	-
Median	-	-	1	-
Q1	2	-	-	-
Rater Effects Not Modeled				
Baseline vs. Composite 1				
Q3	1	-	2	-
Median	1	-	1	-
Q1	2	-	1	-
Baseline vs. Composite 2				
Q3	1	-	3	-
Median	1	-	1	-
Q1	1	-	1	-
Baseline vs. Composite 3				
Q3	-	-	3	-
Median	1	-	1	-
Q1	1	-	-	-
Rater Effects Modeled				
Baseline vs. Composite 1				
Q3	1	-	1	-
Median	-	-	-	-
Q1	-	-	-	-
Baseline vs. Composite 2				
Q3	1	-	-	-
Median	-	-	-	-
Q1	-	-	-	-
Baseline vs. Composite 3				
Q3	1	-	-	-
Median	-	-	-	-
Q1	1	-	-	-

teraction of raters, tasks, and examinee abilities when creating spiraled designs and applying them to assessments that use weighted composite scores. It would also be interesting to obtain results under the spiral design when the number of raters are more than the tasks that have to be rated. Under this condition a semi-nested/spiral design would exist in which a sub-set of students would be rated by some raters while another sub-set would be rated by a different group of raters. The raters within a subset, however, would be spiraled.

This paper should serve as a warning for obtaining composite scores by increasing the weights of OE tasks relative to the MC items for less than substantive reasons, and also for unknowingly pairing extreme raters with extreme-ability-students when the spiraled design is not used, or under the nested design without modeling rater effects.

REFERENCES

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43, 561-563.*

College Board. (1988). The College Board technical manual for the Advanced Placement Program. New York, NY: College Entrance Examination Board.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. Journal of Educational Measurement, 31(1), 56-70.

Engelhard, G. (1994). Examining rater errors in the assessment of written-composition with the many-faceted Rasch model. Jour nal of Educational Measurement, 31(2), 93-112.

Hombo, C. M., Thayer, D. T., & Donoghue, J. R. (2000). A simulation study of the effect of crossed and nested rater designs on ability estimation. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Linacre, J. M. (1989). A user's guide to FACETS, Rasch measure ment program, and Facform, data formatting computer program. Chicago, IL: MESA Press.

Linacre, J. M. (1993). Many-facet Rasch measurement. Chicago, IL: MESA Press.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. Applied Mea surement in Education, 3(4), 331-345.

Taherbhai, H. M. & Young, M. J. (2000). An analysis of rater impact on composite scores using the multifaceted Rasch model. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

JOURNAL OF OUTCOME MEASUREMENT7, 5(1), 839-863
Copyright8 2000, Rehabilitation Foundation, Inc

Measuring disability: application of the Rasch model to Activities of Daily Living (ADL/IADL)

T. Joseph Sheehan, Ph.D.

Laurie M. DeChello

Ramon Garcia

Judith Fifield, Ph.D.

Naomi Rothfield, M.D.

Susan Reisine, Ph.D.

University of Connecticut School of Medicine & University of Connecticut

Requests for reprints should be sent to T. Josep Sheehan, University of Connecticut School of Medicine, 263 Farmington Ave., Farmington, CT 06030

This paper describes a comparative analysis of (ADL) and (IADL) items administered to two samples, 4,430 persons representative of older Americans, and 605 persons representative of patients with rheumatoid arthrisit (RA). Responses are scored separately using both Likert and Rasch measurement models. While Likert scoring seems to provide information similar to Rasch, the descriptive statistics are often contrary if not contradictory, and estimates of reliability from Likert are inflated. The test characteristic curves derived from Rasch are similar despite differences between the levels of disability with the two samples. Correlations of Rasch item calibrations across three samples were .71, .76, and .80. The fit between the items and the samples, indicating the compatibility between the test and subjects, is seen much more clearly with Rasch with more than half of the general population measuring the extremes. Since research on disability depends on measures with known properties, the superiority of Rasch over Likert is evident.

INTRODUCTION

Physical disability is a major variable in health related research. Assessing the degree of difficulty in performing Activities of Daily Living (ADL) and Instrumental Activities of Daily Living (IADL) scales is a common way to measure physical disability. Individuals are asked how much difficulty they have in performing activities of daily living such as dressing, getting up from a chair, or walking two blocks. For each activity, there are 4 possible responses: *no difficulty, some difficulty, much difficulty, or unable to do*. Responses are scored from 1 to 4, or from 0 to 3, summed across all items and averaged to yield a disability score, the higher the average, the greater the disability. These well-validated scales are efficient and widely used but their ordinal scaling can result in distortion by masking ineffective treatments or hiding effective procedures (Merbitz, Morris, & Grip, 1989).

Ordinal scales do not have any obvious unit of measurement, so that addition and division of unknown units is considered meaningless. Wright and Linacre (Wright & Linacre, 1989) have argued that while all observations are ordinal, all measurements must be interval if they are to be treated algebraically, as they are in computing averages. The Rasch (Rasch, 1980) measurement model offers a way to create interval scales from ordinal data, a necessary condition for averaging or analyzing in statistical models that assume interval data.

As Wright and Linacre (Wright & Linacre, 1989) maintain, computing an average score implies a metric that is only available with an interval scale. In addition, those average scores are non-linear and thereby lack a fundamental property assumed in all measurers, from yardsticks to financial worth, that the measures increase along a linear scale. Recognition of this problem is not new. Thorndike (Thorndike, 1904) identified problems inherent in using measurements of this type, such as the inequality of the units counted and the non-linearity of "raw scores". This study demonstrates how the Rasch measurement model converts non-linear ordinal data to a linear, interval scale that provides new information about the utility of commonly used measures of disability. While demonstrating the application of the Rasch model is the main purpose of this study, it also includes a number of comparisons. Rasch person measures are compared to Lickert person scores. Rasch item calibrations are compared to Lickert item scores. Rasch item calibrations estimated on a sample of older Americans are compared to item measures estimated on a sample of American rheumatoid arthritis (RA) patients and to those same items estimated on a sample of British RA patients. These comparisons should enhance understanding of the strengths and perhaps limitations of using the Rasch model in practice.

Background

Before considering Rasch, there is an alternative to averaging ranked data and treating them as interval and that is to simply treat them as ranked data. Responses to each ADL item can be rank ordered, i.e. *no difficulty,* is less than *some difficulty*, is less than *much difficulty,* is less than *unable to do*; so that responses to the ADL tasks can be ordered. Also, the ADL tasks themselves can be ordered. For instance, for most people, *walking two blocks* is more difficult than *lifting a cup or a glass to one's mouth*. It is easy to imagine individuals who, though completely unable to walk two blocks, would have no difficulty lifting a full cup or glass. Because items can be ordered according to a scale of inherent difficulty, ADL items have been organized into hierarchies and disability status is determined by where a person's responses fall along the ordered,

hard-to-easy hierarchy. One such scoring scheme was proposed by Katz (Katz, Downs, Cash, & Grotz, 1970), the creator of the original six item ADL scale. Another step-wise scoring scheme was recently reported by Sonn (Sonn, 1996).

Lazaridis and his colleagues (Lazaridis, Rudberg, Furner, & Casse, 1994) studied the scalability of selected ADL items using criteria associated with Guttman scales. For Guttman (Guttman, 1950), a set of items is a scale if a person with a higher rank than another person is just as high or higher on every item than the other person. Lazaridis found that the Katz scoring scheme fulfilled Guttman's scaling criteria. Lazaridis and his colleagues went further, however, and showed that the Katz hierarchy was one of 360 possible hierarchies, based upon permutations of six ADL items. Lazaridis tested all 360 of these hierarchies using the same Guttman scaling criteria and found four additional scoring schemes that performed equally as well as Katz, and found a total of 103 scoring hierarchies that satisfied minimum standards of scalability according to Guttman.

While Guttman scaling does not violate the ordinal nature of the scales, neither does it produce measures suitable for outcomes analyses that assume interval scaled measures. Also, Guttman's measurement model is deterministic rather than probabilistic and assumes that responses fall along a single hierarchy in a fixed order. The Guttman model would have difficulty with that rare individual who was able to walk two blocks, but unable to lift a full cup to his/her mouth. Daltroy et al. tested a Guttman scale of ADL items to determine whether musculoskeletal function deteriorates in an orderly fashion (Daltroy, Logigian, Iversen, & Liang, 1992). They recommended that *lifting a cup* be dropped because it was too easy. We discuss the item later. Furthermore, the fact that there is not a single hierarchical scale, but as many as 103 different hierarchies underlying Katz' six original ADL items, exposes the disadvantage of a rigid and deterministic hierarchy. A more attractive approach would capture the probabilistic nature of the responses, without losing the concept of a hierarchical scoring function. The Rasch measurement model provides such an alternative.

Rasch, a Danish statistician interested in measuring spelling ability, created a probabilistic measurement function, which simultaneously

estimates the abilities of persons and the difficulty of test items. Rasch showed how the probability of answering a question correctly depended on two things, the ability of the person and the difficulty of the test item. His model estimates person ability and item difficulty simultaneously and shows that the two are independent (Rasch, 1980, p 19). Moreover, the model provides a common scale for assessing both persons and items. The distribution of item difficulties can be examined and compared directly to the distribution of person abilities on the same scale, permitting visual judgments about the appropriateness of these items for these people. Furthermore, the common scale turns out to be an interval scale that is linear in contrast to the non-interval and non-linear scale calculated from the sum of the ranks or the average rank.

METHODS

There are two sets of subjects used in this study. The first set includes 4430 persons who were between ages 50 and 77 during the first National Health And Nutrition Examination Survey (NHANES I), carried out between 1971 and 1975, and who provided complete or nearly complete answers to the 26 ADL items administered during the NHANES Follow-up Study (NHEFS), conducted between 1982 and 1984. There were no initial measures of physical disability administered during NHANES I comparable to the ADL items. NHANES I was administered to a national probability sample of the civilian noninstitutionalized U.S. population (Hubert, Bloch, & Fries, 1993; Miller, 1973).

The persons in the second study group are a nationally representative sample of patients with RA (Reisine & Fifield, 1992). The patients were recruited in 1988, using a two-stage process to ensure that it represented RA patients cared for by board-certified rheumatologists. First, a sample of 116 board-certified rheumatologists was randomly selected from the membership of the American College of Rheumatology. In the second stage, patients with definite or classical RA were asked to participate as they visited the office over a defined recruitment period. Nine hundred twenty-one (88%) of the patients who initially expressed interest agreed to participate in the panel study. Patients were interviewed

yearly by telephone regarding their social, physical and emotional functioning, including their functional ability using the Health Assessment Questionnaire (HAQ) (Fries, Spitz, Kraines, & Holman, 1980). The data for the present analysis are from those patients who participated in the sixth year interview of the panel study (N=605, 66% of the original panel). A recent study (Reisine, Fifield, & Winkelman, 2000) indicates that those who continued to participate had a higher level of education, were more likely to be female, had higher social support, and fewer joint flares.

For comparison of item calibrations, data on a third set of subjects are included, 174 from Great Britain diagnosed with RA (Whalley, Griffiths, & Tennant, 1997). The inclusion of another RA group allows comparison of item calibration between groups of patients with the same diagnosis and between RA patients and the general population of older Americans. Person measures were not available for the British RA group.

The NHEFS data were extracted from the tapes using SAS (SAS Institute, 1989). Initial statistical analyses were performed using SPSS 8.0 (SPSS, 1997) and PRELIS 2.12 (SSI, 1998). Computations for the Rasch model were performed using WINSTEPS (Linacre & Wright, 1998b), a computer program written by Linacre and Wright (Linacre & Wright, 1998a).

Although Rasch created his model with test items which could be scored right or wrong, Andrich (Andrich, 1988) has extended the Rasch model to rating scales which include items with ordered response alternatives such as those found on the ADL scale. Thus each item, instead of being scored right or wrong, is considered to have two or more ordered steps between response categories. The Andrich model estimates the *thresholds* for each item separating each ordered step from the next: that point on a logit scale where a category 1 response changes to a category 2 response, a category 2 response changes to a category 3 response, or a category 3 response changes to a category 4 response. The Andrich model also offers the user a choice between a model that assumes equal steps between categories, the *rating scale model*, or a model that actually estimates the distance between categories, the *partial credit model* (Andrich, 1978), the latter being used in this study to conform to the

Whalley et al. analysis.

The Rasch analysis estimates the difficulty level of each ADL item and the ability level of each person along the same logit scale. The Rasch analysis also produces a test characteristic curve, which graphs the relationship between each person measure and the expected score based on the sum of the ranks, the numerator used to compute a Likert score. In this study, the test characteristic curve for 19 of 26 ADL items from the NHANES study is compared to the test characteristic curve estimated from 19 of 20 parallel items in the Reisine study. The 19 items used in the Reisine study are taken from the twenty item Health Assessment Questionnaire (HAQ) (Fries et al., 1980). The 19 HAQ items contain 15 exact matches and 4 items that are very similar to those used in the NHEFS. The item abbreviations and full text for both ADL and HAQ are shown in Table 1. One of the 26 ADL items, *walk from one room to another on the same level*, had too many missing responses to be included in these analyses. The parallel item from the HAQ, *walk outdoors on flat ground*, was dropped leaving 19 items to compute test characteristic curves for comparison.

RESULTS

The initial analyses showed skewed distributions of responses on all 25 ADL items for the NHEFS sample. The category 1 responses ranged from a high of 95.3% who had no difficulty *lifting a cup*, to a low of 60.0% who had no difficulty *with heavy chores*. The category 4 response, "unable to do" an activity, is uniformly low, under 10% for most items, with *heavy chores* being impossible for 17.4%. A complete table of responses is available from the authors. Skewness is also seen in the responses of the RA patients, although their overall level of disability is higher.

Figure 1 summarizes person disability measures horizontally, and item difficulty level vertically for NHEFS. Persons are distributed across the bottom with M marking the mean and S the standard deviation. There are 102 persons at the mean (M) of -2.27. There are 2,079 persons at the bottom end of the scale, who have no difficulty with any of the ADL

Table 1 Activities of Daily Living (ADL) and Health Assessment Questionnaire items

Item abbreviation	ADL	HAQ
Dresself	Dress yourself, including tying shoes, working zippers and doing buttons	Dress yourself, including tying shoelaces and doing buttons*
Shampoo	Shampoo your hair	Shampoo your hair
Arisechr	Stand up from an armless straight chair	Stand up from an armless straight chair
Inoutbed	Get into and out of bed	Get in and out of bed
Makefood	Prepare your own food	
Cutmeat	Cut your own meat	Cut your meat
Liftcup	Lift a full cup or glass to your mouth	Lift a full glass or cup to your mouth
Openmilk	Open a new milk carton	Open a new milk carton
Wlk2blck	Walk a quarter mile (2 or 3 blocks)	
Wlk2step	Walk up and down at least two steps	Climb up five steps**
Faucets	Turn faucets on or off	Turn faucets on and off
Bathtub	Get in and out of the bathtub	Take a tub bath**
Washbody	Wash and dry your whole body	Wash and dry your entire body
Toilet	Get on and off the toilet	Get on and off the toilet
Combhair	Comb your hair	
Reach5lb	Reach and get down a 5lb. Object (bag of sugar) from just above your head	Reach and get down a 5lb object from just above your head
Pkupclth	Bend down and pick up clothing from the floor	Bend down and pick clothing up from the floor
Cardoors	Open push button car doors	Open car doors**
Openjars	Open jars which have been previously opened	Open jars which have been previously opened
Write	Use a pen or pencil to write with	
Inoutcar	Get in and out of a car	Get in and out of a car
Shop	Run errands and shop	Run errands and shop
Ltchores	Do light chores such as vacuuming	Do chores such as vacuuming and yard work
Liftbag	Lift and carry a full bag of groceries	
Hvchores	Do heavy chores around the house or yard, or washing windows, walls or floors?	

*close match, **modified

items and 15 persons at the right end of the scale who were unable to perform any of the tasks. When extreme persons are included, the mean for persons drops from –2.27 to –3.93, and the standard deviation increases from 1.67 to 2.22. The presence of so many at the bottom of the scale draws attention to the "floor effects" of the test, at least for the general population of older Americans.

The right-hand column in Figure 1 orders items from easiest at the top to hardest at the bottom. Items at the top of the Rasch scale, such as *lift a cup* or *turn a faucet on or off*, are easier than items below, with the hardest items at the bottom, *lift and carry a full bag of groceries* or *do heavy chores around the house or yard*. To the left of each item, the responses, 1, 2, 3, and 4, are arranged at a location corresponding to the expected measure of a person who chose that response to that item. Thus, the expected measure of a person who responded with a 4 to the easiest item, *unable to lift a full cup or glass to one's mouth*, would be slightly greater than 4, or the most disabled end of the scale. Whereas the expected measure of a person who chose category 4 in response to the hardest item and was unable to perform *heavy chores*, would be about -0.8, almost a standard deviation above the mean person disability measure of -2.27.

Figure 1 also shows a colon (:) separating the response categories for each item. Those separation points are the thresholds between response categories estimated by the Andrich model mentioned earlier. The mean item calibration is 0.0 and standard deviation is 1.11. It is noteworthy that the mean item calibration of zero and standard deviation of 1.11 suggests that the item distribution is far to the right of the person distribution. Also, the item standard deviation of 1.11 suggests far less dispersion among items than the dispersion among persons, which reaches 2.22 when extreme measuring persons are included. Such misalignment of the item and person distributions signals serious limitations in using this measure for these persons. The distribution of person measures is far lower than the distribution of item calibrations, indicating a poor fit between this test and these persons, at least at the time these items were administered, that is, the first of several follow-up surveys on the same subjects. For an ideal test, the distribution of items and persons should show similar alignments.

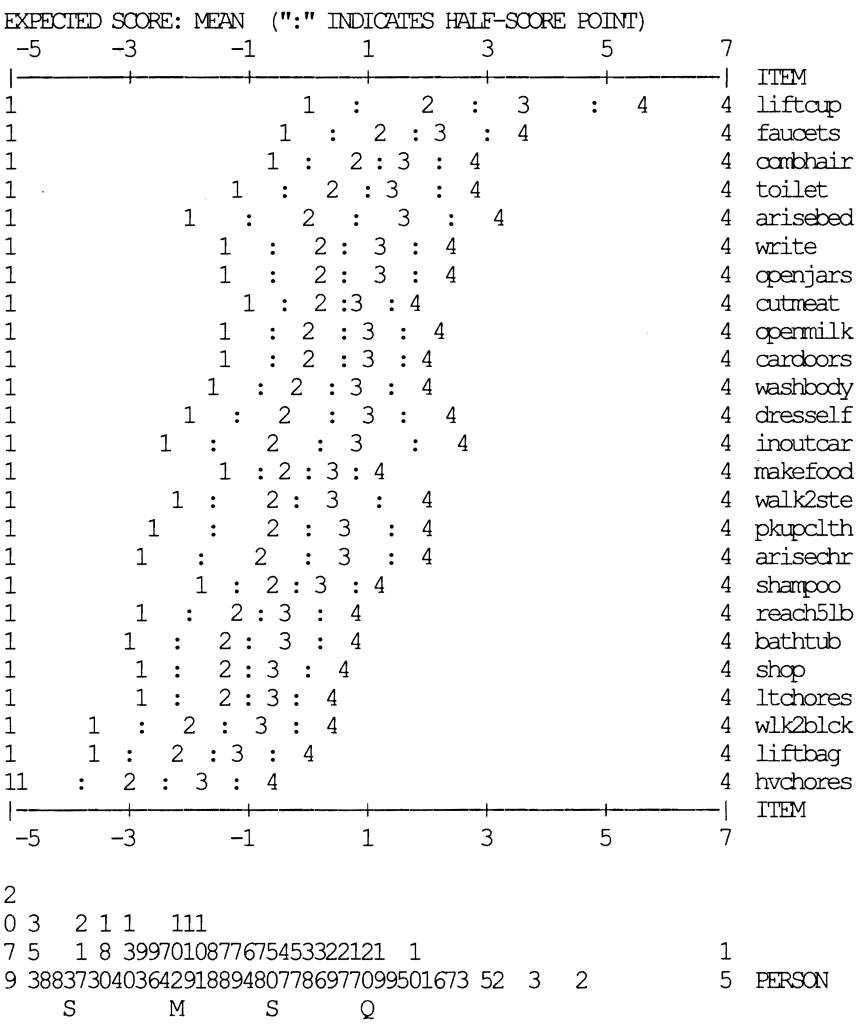


Figure 1. Most probable responses: items are ordered on the right hand side with less difficult items on the top. Persons are ordered on the bottom with more disabled persons on the right. The figure answers the question "which category is a person of a particular person measure most likely to choose?" The distribution of item difficulties has a mean of zero and a standard deviation of 1.04. The distribution of 2310 non-extreme person measures has a mean of -2.18 and a standard deviation of 1.68. There are 2079 extreme responses at the bottom of the scale, and 15 extreme responses at the top of the scale.

The fit may improve as this population ages and becomes more disabled.

In contrast, the Rasch analysis of the RA sample shows a mean of –1.80 for 563 non-extreme persons, indicating a higher level of disability than in the general population. Also, there are only 41 extreme measures, 40 with no difficulty on any item and one person was unable to perform any of the ADL tasks. The misalignment between the item and person distributions is not as severe for the RA patients, but the item distribution, centered about zero with a standard deviation of 0.96, is still higher and more dispersed than the person distribution. It is noteworthy that the mean for non-extreme NHEFS changes little, from –2.27 to –2.12, when 19 rather than 25 ADL items are used; likewise, the mean changes little when extreme persons are included, –3.93 to –3.86.

Figure 2 graphs the relationship between person measures for the 25 ADL items and the Likert score, the average of the sum of the ranks. The graph demonstrates the non-interval and non-linear nature of Likert scores. Thus while the Rasch measures are linear and use the same invariant interval across the entire range of measures, the Likert scores are neither linear nor invariant. Likert metric is about four times greater in the middle than toward the top or bottom of the scale, a clear violation of the linearity assumption underlying all valid measurement. There is a similar curve for the RA sample. While Figure 2 shows the curve relating Rasch measures to observed Likert scores, Winsteps also produces curves for a test based upon statistical expectations, called the test characteristic curves.

Figure 3 shows two test characteristic curves based upon a set of 19 ADL items. One curve is estimated from the NHEFS sample, the second curve from patients with Rheumatoid Arthritis (RA) (Reisine & Fifield, 1992). The NHEFS sample is slightly older on average than the RA sample, 62.0 years versus 59.0 years, and is more male, 43% versus 22%. Although the samples differ slightly in age and considerably in gender and disability with disability levels higher among the RA patients, the test characteristic curves are similar.

While the characteristic curves indicate that the two instruments process raw scores in much the same way, it is helpful to examine the items themselves. Table 2 contains the item calibrations for NHEFS, for the RA sample, and for the RA patients from Great Britain studied by

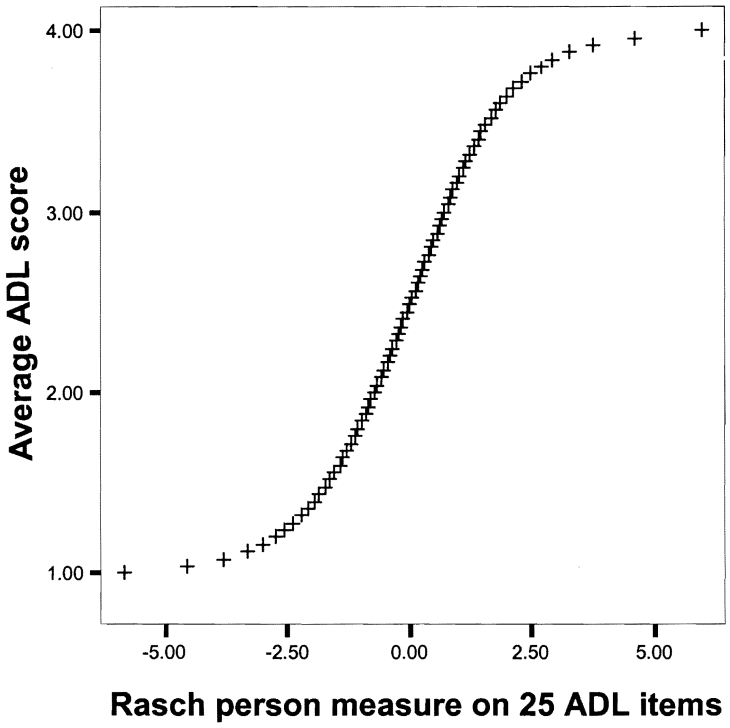


Figure 2. Graph of Rasch measures for persons and ADL average scores based on the average rank assigned to each of 25 ADL items.

Whalley, Griffiths, and Tennant (Whalley et al., 1997). The item error varies from .03 to .08 for NHEFS, from .06 to .09 for HAQ US, and from .11 to .15 for HAQ UK. Ideally, infit root mean square standard errors should be at or near one. *Wash and dry body* was the furthest from one in the NHEFS sample at .70, t*ake a bath* at 1.66 in HAQ US and *wash and dry body* at .58 in HAQ UK were those samples only extreme items. These items do not fit the scales as well as all other items.

Wash and dry body in both the NHEFS and HAQ UK samples may indicate this item is dependent on the data, but it is not so severe for NHEFS. *Take a bath* in the HAQ US sample causes some noise in the scale, however all other items are within .18 of one. It appears that there

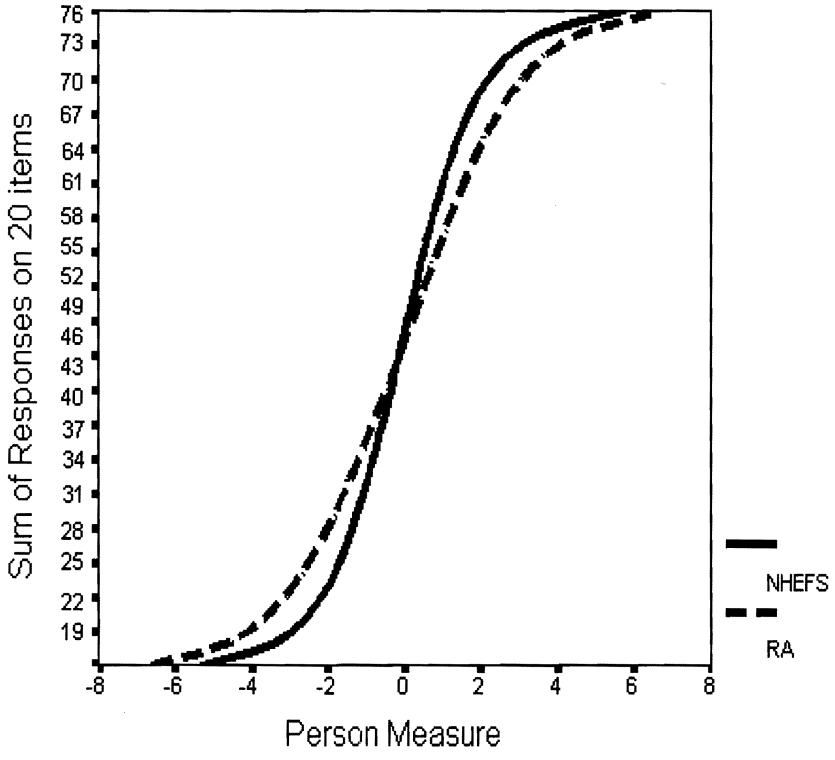


Figure 3. Test characteristic curves for 19 similar ADL items administered to NHEFS and RA samples demonstrating the non-linear nature of scores based upon the sum of the ranks and the similarity of the curves despite differences in the samples.

Table 2 NHEFS, HAQ US, and HAQ UK item comparison. The item measures are from the partial credit Rasch model.

Number	Item	NHEFS Measures	HAQ US Measures	HAQ UK Measures	NHEFS Ranks	HAQ US Ranks	HAQ UK Ranks
1	Liftcup	2.66	.82	2.60	1	5	1
2	Faucets	1.56	.99	.70	2	3	5
3	Toilet	.71	1.53	1.44	3	2	2
4	Arisebed	.56	1.54	.78	4	1	3
5	Openjars	.36	.20	-.66	5	8	14
6	Cutmeat	.29	.11	-.45	6	9	13
7	Openmilk	.29	-1.10	-1.48	7	16	18
8	Cardoors	.25	.30	.11	8	7	8
9	Washbody	.12	.38	-.33	9	6	12
10	Dresself	.02	-.01	.11	10	11	9
11	Inoutcar	-.12	.91	.71	11	4	4
12	Walk2step	-.35	-.14	.10	12	14	10
13	Pkupclth	-.36	.03	.30	13	10	6
14	Arisechr	-.42	-.07	.26	14	12	7
15	Shampoo	-.45	-.13	-.23	15	13	11
16	Reach5lb	-1.15	-1.31	-1.86	16	17	19
17	Bathtub	-1.28	-1.87	-1.09	17	19	17
18	Shop	-1.33	-.35	-.72	18	15	15
19	Ltchores	-1.37	-1.81	-1.01	19	18	16

Item error varies from .03 to .08 for NHEFS, .06 to .09 for HAQ US, and .11 to .15 for HAQ UK. The only extreme infit root mean square errors are *wash and dry body* at .70 and .58 for NHEFS and HAQ UK, respectively, and *take a bath* at 1.66 for HAQ US.

are tasks, which are of different difficulty for RA patients than for the general population of older Americans. Perhaps the most striking difference is the difficulty of *opening a milk carton* where the difficulty for RA patients is –1.10 and –1.48, among their most difficult tasks, as compared to .29 for the general public. It also appears that *getting in and out of cars* is more difficult for the general public than for RA patients: -.12 versus .91 and .71 respectively. Likewise, *getting on and off of a toilet* is easier for RA patients than for the general public: 1.53 and 1.44 versus .71. Perhaps the most striking difference is that of *lifting a full cup or glass to one's mouth* where the American RA patients differ substantially from the British RA patients and from the US general public: .82 versus 2.66 and 2.60 respectively. The size of this difference does not seem spurious because follow-up of US RA patients one and two years later shows that item calibrations for *lifting a cup* were similar: .55 And .85. Daltroy et al. had earlier (1992) recommended replacing *lift a cup* with opening a jar based on an extreme "Rasch Goodness of Fit t" of 7.2. Table 2 shows that opening a jar is more difficult for British RA patients than for the other two groups: -.66 versus .36 and .20 respectively.

Figure 4 shows the two sets of RA items are plotted against each other. The numerals marking each point correspond to the item numbers in Table 2. The correlation between the two sets of items is .80 and if the most discrepant item, *lifting a cup*, is removed, the correlation reaches .87. The correlation between the NHEFS item calibrations and the British RA calibrations are .76, and .71 with the American RA calibrations. The correlation between the NHEFS and US RA items rises from .71 to .77 if the most discrepant item, *lifting a cup*, is removed.

Another way to compare the hierarchical nature of the items is to first rank each item relative to its own sample and then to compare ranks across samples. Where the relative difficulty of *opening a milk carton* was 6th for NHEFS, it was 16th and 18th for the RA samples. Surprisingly, *getting in and out of a car* was 4th for both RA samples, but 11th for NHEFS. *Lifting a cup* was first, or easiest, for NHEFS and the British RA samples, but 5th for the US RA sample. Then there are some American-British differences. *Picking up clothes from the floor* was 13th and 10th for the Americans, while it was 6th for the British. Similarly, *standing up from an armless chair* was 14th and 12th for the Americans, while it was

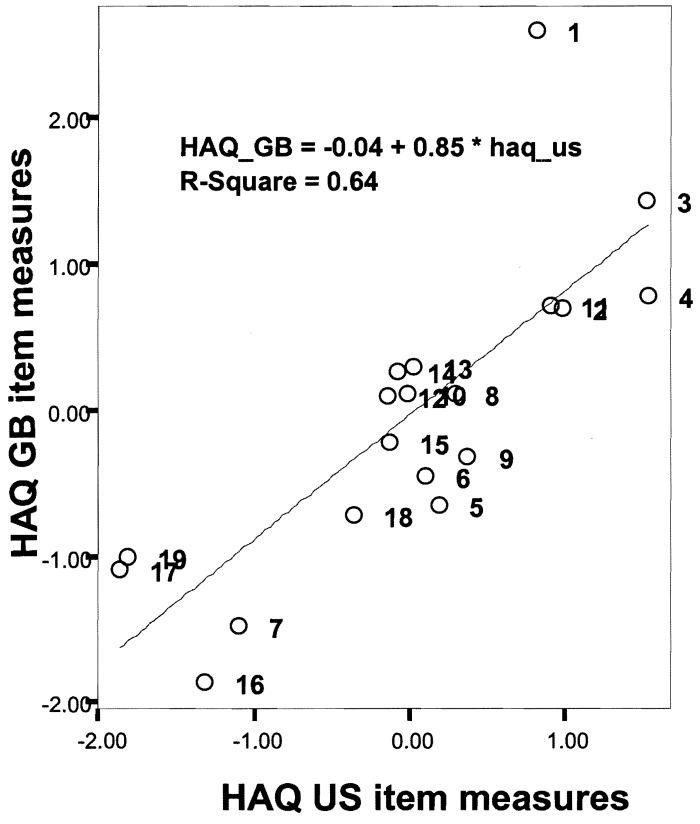


Figure 4. The item measures from the British and US RA samples. Item numbers correspond to Table 2.

7th for the British. Other than these 5 sets of differences, there do not appear to be major differences in the item hierarchical structure among these items. Despite these limitations, there is an overall similarity of hierarchical structure and especially with the similarity of test characteristic

curves, it is interesting to compare the distribution of NHEFS and US RA samples further.

The theoretical independence of person and item estimates was shown early by Rasch (Rasch, 1980) for dichotomous items and its extension to estimating thresholds between categories within each item has already been described by Andrich and van Schoubroeck (Andrich & van Schoubroeck, 1989). Andrich and van Schoubroeck also point out that, "no assumptions need to be made about the distribution of β_n in the population (Andrich & van Schoubroeck, 1989, p. 474)." As shown below the person distributions are dramatically different for the NHEFS and RA samples.

Figure 5 shows the distribution of persons and items for the RA and NHEFS samples based on Likert scoring. To see how sample dependent the item calibrations are for the Likert scoring, compare the distribution of item averages in Figure 5, where they are much more bunched together for the NHEFS sample than they are for the RA sample. The pattern of item calibrations for Rasch measurement is more similar across samples as seen in Figure 6. It is clear from both figures that the distribution of disability is different in RA patients than in the general population. Both figures show that the NHEFS sample has a much higher percentage of persons at the lower end of the scale. Both Likert and Rasch also show the reversed J shape for the NHEFS sample, and a flatter distribution of RA patients. It would appear, at least at first, that both Likert scores and Rasch calibrations provide similar information about the distribution of disability in these two populations. Important differences become apparent when the moment statistics, mean, standard deviation, skewness, and kurtosis shown in Table 3, enhance the information available from the graphical images.

The means show a higher level of disability in the RA sample, but the differences seem greater on the Rasch scale. The Likert means appear closer, 1.24 and 1.75, than the Rasch means, –3.87 and –2.10, although they are both about one standard deviation apart in either metric. The standard deviations seem to show a much greater spread on the Rasch scale, but are similar for both the Likert scoring and Rasch measurement for the two samples. While the means and standard deviations appear to

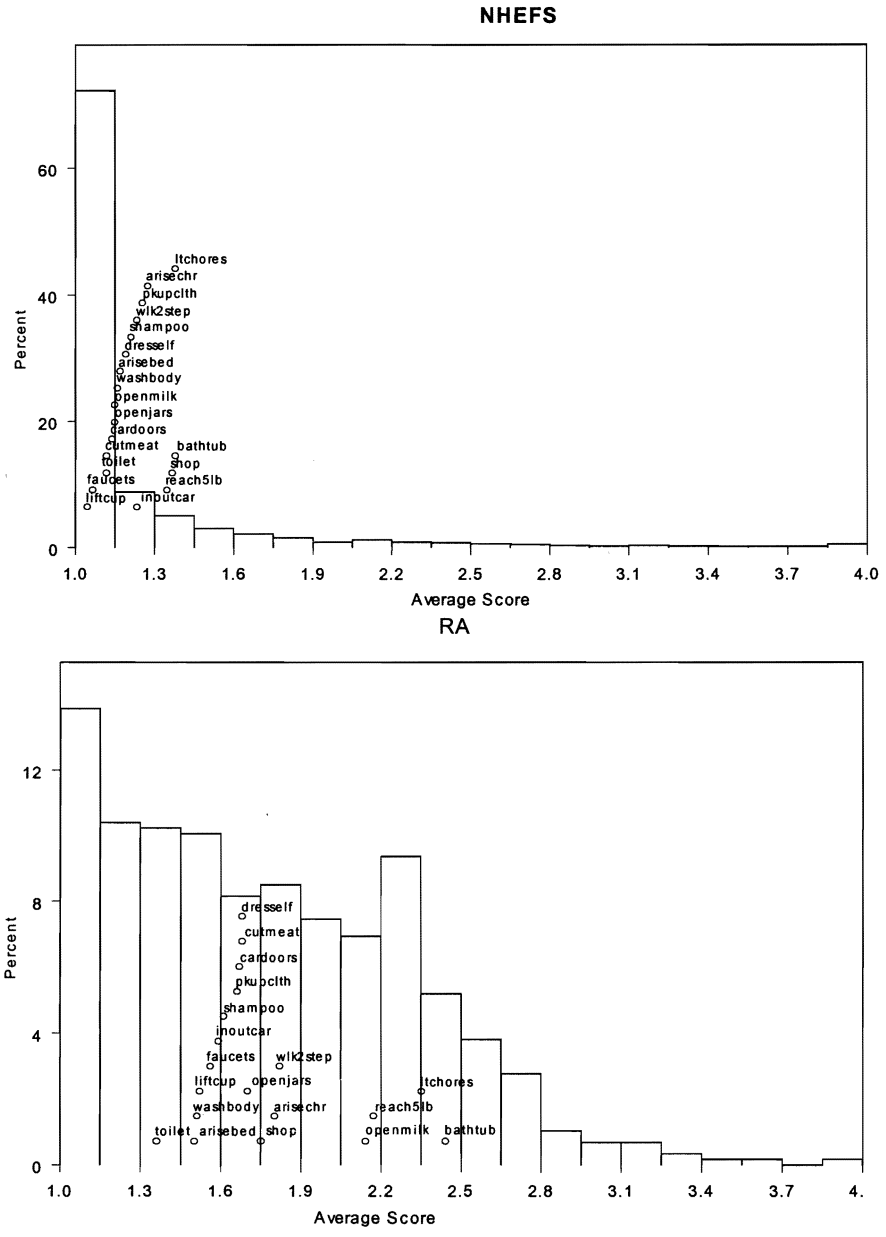


Figure 5. Distribution of Likert scores for persons and items based on 19 common items, for NHEFS and RA samples. Means = 1.24 and 1.75, standard deviations = .52 and .54, skewness = 2.95 and .75, and kurtosis = 9.22 and .72.

be different, they are at least consistent in describing the two samples.

The skewness indices reflect the asymmetry of the distributions. The NHEFS skewness index for the Likert scores is nearly twice as large as the index for Rasch measures, 2.95 versus 1.39. A normal distribution has a skewness index of zero and a positive index describes a long tail to the right. The skewness indices for the RA sample show a reversal of signs, 0.75 to -0.45. With skewness, therefore, consistency is lost and the information conveyed by Rasch is both contrary and contradictory to information conveyed by the Likert scores.

The indices of kurtosis, describing how peaked or flat the distributions are, show almost a 5 fold difference between Likert and Rasch for the NHEFS sample, 9.22 and 1.89, while the indices are close for the RA sample, .60 for Rasch and .72 for Likert. Since normal theory statistics assume there are excesses neither in skewness nor kurtosis, it is clear that the distributional properties of the Rasch measures are superior to those of Likert. The Likert scores provide information about skewness and kurtosis that is contrary if not contradictory to the information provided by the Rasch analysis. The fact that the correlation between Likert scores and Rasch measures is extremely high, obviously, does not mean the measures are equivalent; it simply means that persons are in the same rank order on both scales. Furthermore, high correlations distract from the real need of a meaningful metric against which progress can be measured.

Finally, it is noteworthy that the Rasch reliability estimates are consistently lower than reliabilities estimated from Likert scoring. The Rasch reliability of the person measures estimated from 25 ADL items is .86 for non-extreme persons, and drops to .62 when the 2094 extreme person measures are included. The person measure reliability for the RA patients based upon 19 HAQ items is .90 and drops slightly to .88 when 41 extreme person measures are included. The coefficient alpha estimate of reliability for these same items when scored as a Likert scale is .94. In the Rasch analysis, the HAQ has greater precision and less measurement error in assessing patients with arthritis than the ADL has in the general population, and both have more measurement error associated with person measures than suggested by coefficient alpha from a Likert scale.

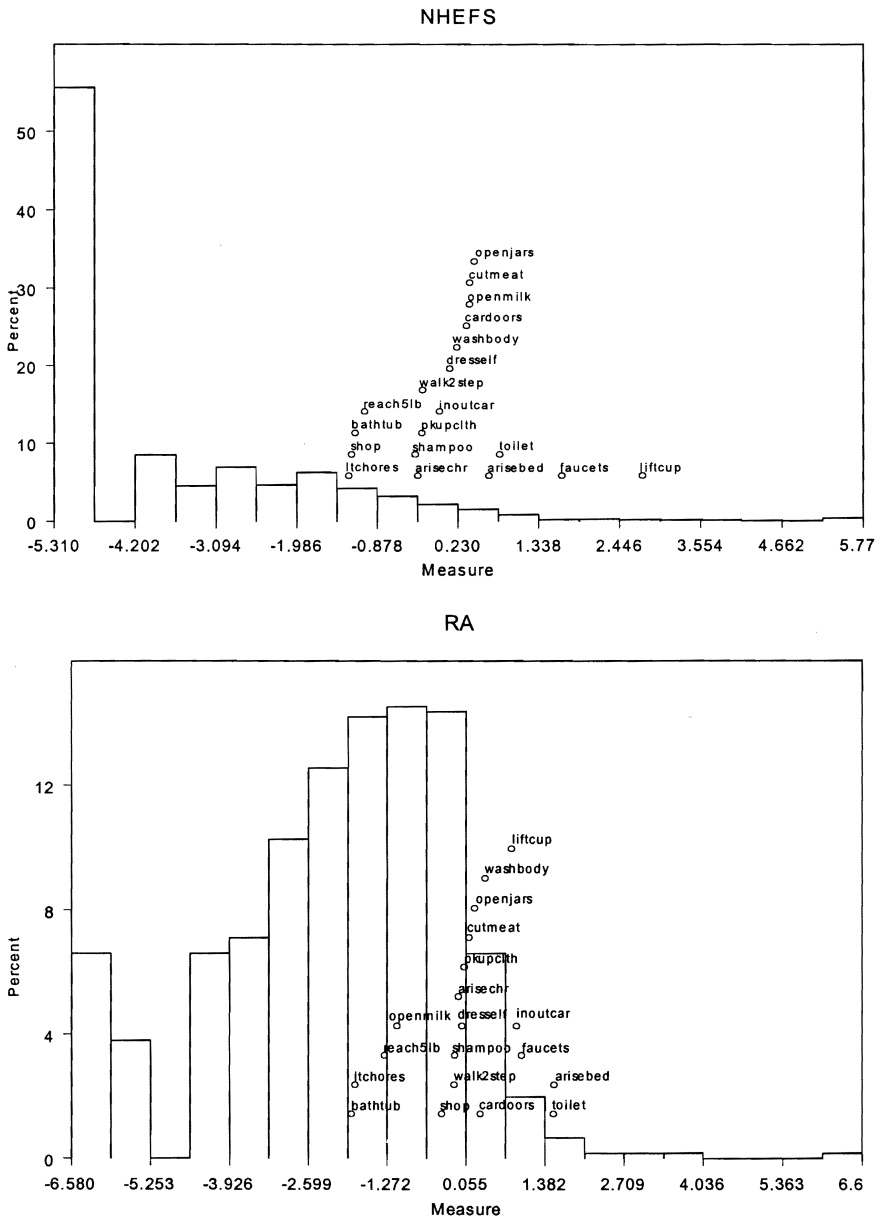


Figure 6. Distribution of Rasch measures for NHEFS and RA samples and items. Means = -3.87 and -2.10, standard deviations = 1.97 for both, skewness =1.39 and -0.45, and kurtosis = 1.89 and .60.

Table 3. Summary statistics to describe distribution of disability in NHEFS and RA samples using Likert scores and Rasch measures: Means, Standard Deviation, Skewness (measure of symmetry and equals 0.0 for normal distribution; tail to the right for posivite value, and left for negative value), and Kurtosis (measure of peakness: equals 0.0 for normal distribution).

Scale	Sample	Mean	SD	Skewness	Kurtosis
Likert	NHEFS	1.24	0.52	2.95	9.22
	RA	1.75	0.54	0.75	0.72
Rasch	NHEFS	-3.87	1.97	1.39	1.89
	RA	-2.10	1.97	-0.45	0.60

CONCLUSIONS

Improved precision in the methods of measuring disability is essential to disability research. Thurstone set out the fundamentals of measurement in the 1920s (Wright, 1997) and described the universal characteristics of all measurement. Measurements must be *unidimensional* and describe only one attribute of what is measured. The Rasch model assumes a single dimension underlying the test items and provides measures of fit, so that the unidimensionality assumption can be examined. Measurements must be *linear* because the act of measuring assumes a linear continuum as for example, length, height, weight, price, or volume. The Rasch analysis

reveals the non-linear nature of Likert scores. Measurements must also be invariant and use a repeatable metric all along the measurement continuum, the essence of an interval scale. The Rasch analysis reveals the lack of a repeatable metric for Likert scores.

While all ADL items showed adequate fit to the Rasch model, and hence the unidimensionality requirement has been met, further study may show that the single construct of disability may have more than one measurable feature. In fact, some HAQ users combine items into up to 8 subscales (Tennant, Hillman, Fear, Pickering, & Chamberlain, 1996), and report similarity in overall measures using either subscales or measures based on all of the items. They also report large logit differences among items within a subscale. Daltroy et al. (Daltroy et al., 1992) grouped items into six subscales in search of a sequential functional loss scale. Their aim was to develop a Guttman scale of function in the elderly. They reported that 83% of non-arthritic subjects fit their scale compared to 65% of arthritic subjects. Their Rasch analysis indicated that *lifting a cup to the mouth* did not fit the Guttman scale pattern. In the Daltroy study, *getting in and out of a car*, was grouped with *doing chores* and *running errands* as the most difficult subscale. In the current study, *getting in and out of a car* was easier for US RA patients, with a rank of 4th easiest compared to *lifting a cup*, which was 5th. It was also 4th easiest for British RA patients and 11th easiest for the NHEFS sample. The current study along with these findings may signal cautions conceptualizing disability as a single invariant sequence or hierarchy. The Rasch assumption of unidimentionality seems intact and we have conducted confirmatory factor analysis using methods suitable for ordinal items and found that a single factor model adequately fit the ADL data. Nonetheless, the nature of disability may well vary within and between groups that even share the same diagnosis, such as RA. Such variation may be seen in the item calibrations for *lifting a cup*.

The Rasch analysis also demonstrates that the ADL based measures of disability provide a poor measurement tool for nearly half of the NHEFS sample who were entirely disability free during the 1982 to 1984 follow-up. Certainly, the fit between items and persons is better in a population with more disability such as RA patients. However, even for RA patients,

the estimate of reliability for the Likert scale is inflated over that provided by the Rasch analysis, and the information provided by the Likert analysis about the distribution of disability in either the general population or among RA patients is different from and contradictory to the information provided by the Rasch analysis.

ACKNOWLEDGEMENTS

Supported in part by grants from the Arthritis Foundation and the Claude Pepper Older Americans Independence Center. We thank Dr. Alan Tennant from the University of Leeds for his prompt response to our request for HAQ item data for RA patients from the UK.

REFERENCES

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Andrich, D. (1988). *Rasch Models for Measurement.* Newbury Park, CA: Sage Publications, Inc.

Andrich, D., & van Schoubroeck, L. (1989). The General Health Questionnaire: a psychometric analysis using latent trait theory. *Psychol Med, 19*(2), 469-85.

Daltroy, L. H., Logigian, M., Iversen, M. D., & Liang, M. H. (1992). Does musculoskeletal function deteriorate in a predictable sequence in the elderly? *Arthritis Care Res, 5*(3), 146-50.

Fries, J. F., Spitz, P., Kraines, R. G., & Holman, H. R. (1980). Measurement of patient outcome in arthritis. *Arthritis and Rheumatism, 23*(2), 137-145.

Guttman, L. (1950). The basis for scalogram analysis. In Stouffer (Ed.), *Measurement and Prediction* (Vol. 4, pp. 60-90). Princeton, N.J.: Princeton University Press.

Hubert, H. B., Bloch, D. A., & Fries, J. F. (1993). Risk factors for physical disability in an aging cohort: the NHANES I Epidemiologic Followup Study [published erratum appears in J Rheumatol 1994 Jan;21(1):177]. *J Rheumatol, 20*(3), 480-8.

Katz, S., Downs, T. D., Cash, H. R., & Grotz, R. C. (1970). Progress in development of the index of ADL. *Gerontologist, 10*(1), 20-30.

Lazaridis, E. N., Rudberg, M. A., Furner, S. E., & Casse, C. K. (1994). Do activities of daily living have a hierarchical structure? An analysis using the longitudinal study of aging. *Journal of Gerontology, 49*(2, M47-M51), M47-M51.

Linacre, J. M., & Wright, B. D. (1998a). *A User's Guide to Bigsteps Winsteps: Rasch-Model Computer Program*. Chicago: MESA Press.

Linacre, J. M., & Wright, B. D. (1998b). Winsteps. Chicago: MESA Press.

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference [see comments]. *Arch Phys Med Rehabil, 70*(4), 308-12.

Miller, H. W. (1973). Plan and operation of the Health and Nutrition Examination Survey: United States- 1971-1973. *Vital and Health Statistics, Series 1*(10a), 1-42.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: MESA Press.

Reisine, S., & Fifield, J. (1992). Expanding the definition of disability: implications for planning, policy and research. *Milbank Memorial*

Quarterly, 70(3), 491-509.

Reisine, S., Fifield, J., & Winkelman, D. K. (2000). Characteristics of rheumatoid arthritis patients: who participates in long-term research and who drops out? *Arthritis Care Res, 13*(1), 3-10.

SAS Institute (1989). SAS (Version 6.0). Cary, NC: SAS Institute Inc.

Sonn, U. (1996). Longitudinal studies of dependence in daily life activities among elderly persons. *Scandinavian Journal of Rehabilitation Medicine, S34*, 2-28.

SPSS. (1997). SPSS (Version 8.0). Chicago: SPSS.

SSI. (1998). PRELIS (Version 2.20). Lincolnwood, IL: Scientific Software International.

Tennant, A., Hillman, M., Fear, J., Pickering, A., & Chamberlain, M. A. (1996). Are we making the most of the Stanford Health Assessment Questionnaire? *British Journal of Rheumatology, 35*, 574-578.

Thorndike, E. L. (1904). *An Introduction to the Theory of Mental and Social Measurements*. New York: Teacher's College.

Whalley, D., Griffiths, B., & Tennant, A. (1997). The Stanford Health Assessment Questionnaire: a comparison of differential item functioning and responsiveness in the 8- and 20 item scoring systems. *Brit J Rheum, 36*(Supple 1), 148.

Wright, B. (1997). *A history of social science measurement*. MESA Memo 62. Available: http://MESA.spc.uchicago.edu/memo62.htm [1998, Oct. 29, 1998].

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*, 857-867.