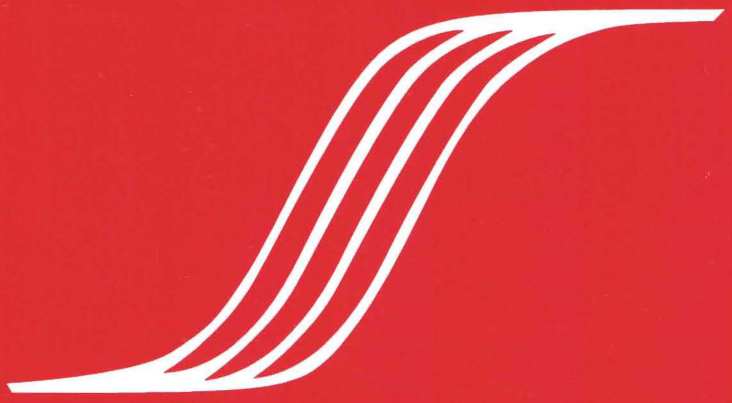
Volume 3, Number 4, 1999 ISSN 1090-655X

Journal of Outcome Measurement®

Dedicated to Health, Education, and Social Science



REHABILITATION FOUNDATION INC.
EST. 1993
Research & Education

Editor

Richard M. Smith ..... Rehabilitation Foundation, Inc.

Associate Editors

Benjamin D. Wright ..... University of Chicago
Richard F. Harvey ..... RMC/Marianjoy RehabLink
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
Allen Heinemann ..... Rehabilitation Institute of Chicago
Mark Johnston ..... Kessler Institute for Rehabilitation
David McArthur ..... UCLA School of Public Health
Robert Rondinelli ..... University of Kansas Medical Center
Tom Rudy ..... University of Pittsburgh
Mary Segal ..... Moss Rehabilitation
Alan Tennant ..... University of Leeds
Luigi Tesio ..... Foundazione Salvatore Maugeri, Pavia
Craig Velozo ..... University of Illinois Chicago

Educational/Psychological Editorial Board

David Andrich ..... Murdoch University
Trevor Bond ..... James Cook University
Ayres D'Costa ..... Ohio State University
Barbara Dodd ..... University of Texas, Austin
George Engelhard, Jr. ..... Emory University
Tom Haladyna ..... Arizona State University West
Robert Hess ..... Arizona State University West
William Koch ..... University of Texas, Austin
Joanne Lenke ..... Psychological Corporation
J. Michael Linacre ..... MESA Press
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 3, Number 4 1999

Articles

Indexing/Abstracting Services: JOM is currently indexed in the *Current Index to Journals in Education* (ERIC), *Index Medicus*, and MEDLINE. The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Editor's Note

Journal of Outcome Measurement is pleased to present the first installment in a new series of articles which focus on understanding the underlying concepts of Rasch measurement, appearing under the general title "Understanding Rasch Measurement". Beginning with this issue and extending over the next five years, the journal will publish one article per issue that looks at an issue, methodology, or technique in Rasch measurement from an instructional or informational point of view. The purpose of this series is to give the readership a fuller understanding of the fundamental principals that underlie many of the methodologies used in research articles published in this journal. The editorial board has contacted many of the leading methodologists in Rasch measurement and they have agreed to write chapters dealing with their special areas of expertise. These chapters, taken together, will create an introduction to Rasch measurement, and are presented in an order that will approximate the steps in acquiring Rasch measurement skills.

The first installment, contained in this issue, contains a review of the various procedures that are used to estimate the variety of parameters available in the family of Rasch measurement models. Michael Linacre's discussion and comparison of estimation procedures gets the series off to a rousing start. The first issue of Volume 4, due out in January 2000, will contain a introduction to Rasch measurement models and a discussion of the interrelationship between these measurement models written by Benjamin D. Wright and Magdalena Mok. Future topics include issues in the analysis of fit, metric development and score reporting, four articles discussing methods of analyzing different types of data, item banking and test equating, bias analyses, standard setting, performance based assessment, and many more. We hope that you will find this series useful and welcome suggestions from readers for new topics to include in the series.

Reviewer Acknowledgement

The Editor would like to thank the following people who provided manuscript reviews for the Journal of Outcome Measurement, Volume 3.

David Andrich—Murdoch University, Australia

Betty Bergstrom—CAT Inc., Evanston, IL

Rita Bode—Rehabilitation Institute of Chicago

Trevor Bond—James Cook University, Australia

Karon Cook, Rehab R&D Center, Houston VA Medical Center

Ayres D'Costa—Ohio State University

Barbara Dodd—University of Texas at Austin

Graham Douglas—University of Western Asutralia

George Engelhard, Jr.—Emory University

William Fisher, Jr.—Louisiana State University Medical Center

Thomas Haladyna—Arizona State University West

Robert Hess—Arizona State University West

George Karabatsos—Louisiana State University Medical Center

William Koch—University of Texas at Austin

Gene Kramer—American Dental Association

Anna Kubiak—Educational Testing Service

Joanne Lenke—The Psychological Corporation

J. Michael Linacre—University of Chicago

David McArthur—UCLA School of Public Health

Carol Myford—Educational Testing Service

Nambury Raju—Illinois Institute of Technology

Randall Schumacker—University of North Texas

Everett Smith—University of Illinois at Chicago

Alan Tennant—University of Leeds, England

Craig Velozo—University of Illinois at Chicago

Benjamin D. Wright—University of Chicago

David Zurakowski—Children's Hospital, Boston

JOURNAL OF OUTCOME MEASUREMENT®, *3*(4), 297–307
Copyright© 1999, Rehabilitation Foundation, Inc.

A Validation Study of the Daily Activities Questionnaire: An Activities of Daily Living Assessment for People with Alzheimer's Disease

Frances Oakley
National Institutes of Health, Bethesda, MD

Jin-Shei Lai
University of Illinois at Chicago, Chicago, IL

Trey Sunderland
National Institutes of Health, Bethesda, MD

The Daily Activities Questionnaire (DAQ) was developed to assess activities of daily living (ADL) independence in people with Alzheimer's disease. After administering it to 276 people diagnosed with Alzheimer's disease, we examined the quality of the rating scale and its structure using a Rasch measurement approach. Results indicated that the original 10-point rating scale should be restructured to a 5-point rating scale to improve the quality of the instrument. In addition, we found that all but two ADL items defined the same construct and could be combined into a single summary measure of ADL independence. The remaining items were positioned along a hierarchical continuum, with IADL tasks more difficult than PADL tasks. Furthermore, the tasks were logically ordered by difficulty. We therefore report that the DAQ is a valid scale and conclude that it is a viable measure of ADL independence for studies of Alzheimer's disease.

Requests for reprints should be sent to Frances Oakley, National Institutes of Health, Occupational Therapy Section, Warren G. Magnuson Clinical Center, Building 10, Room 6S235, 10 Center DR MSC 1604, Bethesda, MD 20892-1604, e-mail: Fran_Oakley@nih.gov.

Introduction

Progressive changes in cognition, behavior, and independence in performing activities of daily living (ADL) characterize Alzheimer's disease. Knowing the ADL ability of a person with Alzheimer's disease serves as a basis for designing interventions that provide sufficient help without offering unnecessary assistance that may reduce self esteem or increase the burden of care on the caregiver. Those who care for people with Alzheimer's disease need to be able to accurately identify persons who experience difficulty when performing ADL. Reliable and valid measures of ADL independence in people with Alzheimer's disease are essential in both patient care and clinical research for evaluating outcomes, planning placement, estimating care requirements, and indicating change in ADL status.

The Daily Activities Questionnaire (Oakley, et al., 1991) was developed to assess independence in personal (PADL) and instrumental activities of daily living (IADL) of people with Alzheimer's disease. The interrater reliability of the Daily Activities Questionnaire was examined in a pilot study. Excluding two items (walking and recreation that were non-significant), item by item analysis revealed intraclass correlation coefficients, significant at the $p<0.05$ level, ranging from 0.60 to 0.80 (Oakley, et al., 1991). However, no study has been done to examine the internal and construct validity of the questionnaire. Internal validity in this context addresses whether the instrument measures what it purports to measure, while construct validity tests if the item hierarchy fits the theoretical model upon which the questionnaire was based. Without verifying internal and construct validity, we cannot assume the results from the questionnaire are valid.

The purpose of this study was to examine the quality of the Daily Activities Questionnaire (DAQ) and its structure. Two research questions were posited:

1) Do the 14 DAQ items work together to define a single construct of ADL independence?
2) Do the items define a theoretical linear continuum of increasing difficulty?

To answer these questions, we conducted a retrospective review of our experience with the DAQ, applying the technique of Rasch analysis (FACETS, Linacre, 1996) to our data.

Methodology

Subjects

The sample of convenience included 276 persons (121 M, 155 F, 268 Caucasians, 8 African Americans) age range 50 to 87 (Mean age=68.2, SD=8.0) diagnosed with Alzheimer's disease who were hospitalized on the Geriatric Psychiatry Branch at the National Institutes of Mental Health, National Institutes of Health (NIH) between 1986 and 1996. All subjects were clinically diagnosed with dementia based on the Diagnostic and Statistical Manual of Mental Disorders, 3rd ed., Revised (DSM-III-R; American Psychiatric Association, 1987) and met the criteria of the National Institute of Neurological and Communicative Disorders and Stroke—Alzheimer's Disease and Related Disorders Association for probable Alzheimer's disease (McKhann, et al., 1984). To achieve a homogenous sample of people with Alzheimer's disease, subjects were excluded if they had other mental or physical impairments such as stroke, cancer, or Parkinson's disease.

Procedure

The DAQ is a 14-item written questionnaire designed to assess independence in PADL and IADL. The PADL items included in the questionnaire are: bathing, dressing, toileting, walking, eating, and grooming. The IADL items are home care, cooking, shopping, finances, and phone use. Additional items include sleeping, recreation, and a rating for "overall independence." A visual analog scale comprising a 99-mm line with the lowest score associated with greater dependence represents each item. The PADL, IADL, and "overall independence" scales range from "totally dependent" to "totally independent," the sleeping scale ranges from "unable to follow a normal sleep schedule" to "sleep schedule is normal," and the recreation scale ranges from "no longer pursues recreation" to "participates in recreation like old self." After directly observing subjects perform ADL (excluding sleeping) over a three-week, inpatient admission, an occupational therapist rated each subject using the questionnaire. The item "sleeping" was rated based on the nurse report of the subject's ability to follow a normal sleep schedule. We then equally divided the 0 to 99 mm lines into 10 categories (Linacre, 1998) and converted them to a 10-point rating scale based on the actual length of the line a rater marked from the totally dependent (left) end of the line. For example, a score of 5 was assigned if a rater marked between 50 and 59 mm on the line,

while a score of 9 was assigned to an item if the mark fell between 90 and 99 mm. We did not include the "overall independence" item in the analysis since it represents the summary of the whole instrument. Only one rater (i.e., the first author) rated the DAQ in this study. Therefore, no additional rater severity adjustment was conducted.

Data Analysis

We used the Rasch measurement model FACETS (Linacre, 1996) to analyze the data generated from the DAQ. When the data fit the model, Rasch analysis converts the ordinal raw scores from the 10-point rating scale into interval measures expressed as log-odds units or logits. Rasch analysis determines internal validity by asking if the items of the instrument meet the criteria of unidimensionality. Unidimensionality refers to whether an instrument measures a single dominant construct, even though multiple attributes are measured. The DAQ was designed to measure independence in daily activities and includes items representing elements that are theorized to be components of activities of daily living. Unidimensionality is a necessary precursor to combining items to obtain a total score for the questionnaire. If items were found not to measure the same domain construct, results that include the misfitting items would yield misleading information.

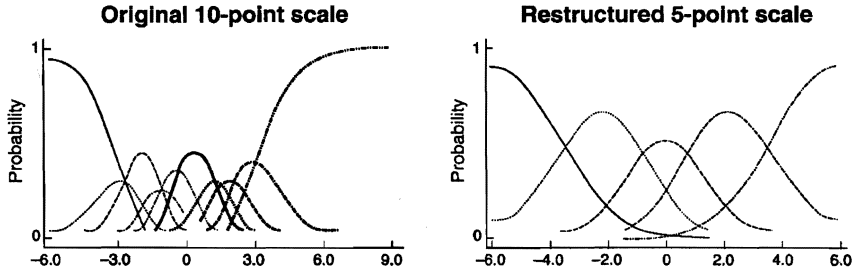
Rasch analysis determines the unidimensionality of the instrument by examining *response patterns* of the items in the instrument, which are demonstrated by *infit* Mean Square (MnSq) statistics. The MnSq is the ratio between observed and expected variance (Wright and Masters, 1982). The expected MnSq value is 1.00, which indicates observed variance is the same as expected variance. However, limited random variance is allowed. An acceptable range for an *infit* MnSq value for a rating scale is between 0.6 and 1.4 (Wright and Linacre, 1994). An item with a MnSq value smaller than 0.6 indicates that the item does not provide additional information beyond the rest of items on the scale. Nonetheless, such an item still defines the same construct as the rest of the items even though it does not improve the measurement quality of the instrument. If the MnSq value for an item is greater than 1.4, either the item does not define the same construct as the rest of the items in the instrument or it is ambiguously defined.

Rasch analysis also provides a significance test for the MnSq value termed the ZStd. A ZStd value greater than 2.0 indicates that the corresponding MnSq value is significant at the 0.05 level. Accordingly, the misfit criterion for this study was MnSq > 1.4 along with a ZStd > 2.0.

Prior to calculating fit statistics, we examined the rating scale structure to ensure results using this rating scale are valid. If a rating scale is valid, ordered step calibrations should be discovered. The step calibration is the calibrated difficulty of *this* step (the transition from the category below to *this* category) relative to the prior step (Linacre, 1998). The bottom step has no prior step, and so the difficulty is shown as NONE. The step calibration is expected to increase with category value since less able persons are more likely to obtain lower categories while more able persons are more likely to obtain higher categories. For this study, a 10-point rating scale (i.e., 0 = scores 0 to 9, 1 = scores 10 to 19, 2 = scores 20 to 29, and so forth) would expect to yield 9 ordered steps. Disordered steps indicate that the instrument raters do not comprehend the rating scale clearly, and thereby do not use the rating scale as expected. The rating scale needs to be re-structured if disordered steps are found.

Results

The results of the rating scale structure are shown in Figure 1. We expected that we would have 9 ordered step measures for our 10-point



Category	Step Calibrations Measure	S.E.	Category Counts
(0-9 mm)			18
(10-19 mm)	-2.73*	.26	18
(20-29 mm)	-2.78*	.23	14
(30-39 mm)	-1.11*	.20	12
(40-49 mm)	-1.12*	.20	17
(50-59 mm)	-0.35	.20	16
(60-69 mm)	1.34	.24	26
(70-79 mm)	1.53	.29	22
(80-89 mm)	2.09	.36	25
(90-99 mm)	3.14	.50	100

*Disordered steps

Category	Step Calibrations Measure	S.E.	Category Counts
(0-9 mm)			563
(10-39 mm)	-3.35	.08	615
(40-59 mm)	-0.77	.07	550
(60-89 mm)	0.72	.07	794
(90-99 mm)	3.40	.06	941

Figure 1. Category probability curves and corresponding step calibrations for the original 10-point and restructured 5-point DAQ rating scales.

Table 1

Comparison of the Instrument Quality Indicators

	Analysis 1	Analysis 2	Analysis 3
Numbers of items	12[a]	12	12
Rating scale	5-point (center=4)[b] 0, (1-3), 4, (5-8), 9	5-point (center =5) 0, (1-4), 5, (6-8), 9	5-point (center=4&5) 0, (1-3), (4-5), (6-8), 9
Step calibrations	Disordered (-3.36[c], -.21, -.27, 3.84)	Disordered (-3.74, .66, -.08, 3.16)	Ordered (-3.35, -.77, .72, 3.40)
Misfitting item	Sleeping (2.9)[d]	Walking (1.5) Sleeping (3.2)	Walking (1.4) Sleeping (3.1)
Item separation	17.35	17.53	17.14
Person separation	4.00	3.90	4.01

a. overall item was not included in the analysis
b. indicate initial ratings: 0=scores 0 to 9, 1=10 to 19, 2=20 to 29, and so forth. The new rating scale was formed via combining initial ratings 1, 2 and 3 as well as 5, 6, 7 and 8.
c. indicate calibrations of step 1 to 2, 2 to 3, 3 to 4, and 4 to 5 respectively.
d. indicates *infit* MnSq value.

rating scale. However, we found disordered steps for category 1 to 2 and 3 to 4. Also, although step calibrations for category 5 to 6 and 6 to 7 were from lower to higher, almost no significant differences among these calibrations could be found (step measures were $1.34 \pm .24$ and $1.53 \pm .29$ respectively). These results suggested that our 10-point rating scale could not discriminate patients' ADL independence in a consistent manner. Therefore, to improve our scale, we restructured it. As Nunnally (1967) has recommended, this 10-point rating scale was converted into a 5-point rating scale to get better results. Table 1 summarizes why this 5-point scale was used. The corresponding probability curves of the final rating scale (i.e., combining scores 10 to 39, 40 to 59, and 60 to 89) used for this study is shown in Figure 1. After different trials, a 5-point rating scale that combined 1 to 3, 4 and 5, and 6 to 8, showed the best structure and was used for further analysis. No disordered step measures were detected in our revised scale, and each category could be easily distinguished from every other category (shown in Figure 1).

The fit statistics were calculated by Rasch analysis. Initial analysis showed that one item, sleeping (*infit* MnSq = 3.1, ZStd = 9.0) misfit according to the criteria of the fit statistics. This result was expected since sleeping is not typically considered a PADL or an IADL. Sleeping was included in the questionnaire because families reported that disturbed sleep patterns often led to nursing home placement (Oakley et al., 1991). Another analysis was conducted with the "sleeping" item removed. Results (Table 2) showed that "walking" misfit (*infit* MnSq = 1.7, ZStd = 5.0).

Table 2

Summary of Rasch Analyzed Results (Items arranged by item measure or endorsability)

		Measure	Real S.E.	Infit MnSq	Infit ZStd	
(Hardest item)	Finances—IADL	4.91	.14	0.9	-1	
	Shopping—IADL	2.62	.16	0.5	-6	
	Cooking—IADL	2.57	.16	0.5	-6	
	Home care—IADL	2.01	.15	0.5	-5	
	Recreation	1.75	.12	1.3	3	
	Phone use —IADL	1.55	.11	1.0	0	
	Grooming—PADL	-1.17	.16	0.5	-6	
	Dressing—PADL	-1.49	.15	0.6	-4	
	Bathing—PADL	-1.66	.16	0.6	-5	
	Walking—PADL	**-3.48**	**.17**	**1.7**	**5...misfit**	
	Toileting—PADL	-3.56	.14	1.2	2	
(Easiest item)	Eating—PADL	-4.04	.15	0.8	-1	
Mean (Count: 12)		.00	.15	0.8	-2.3	
S.D.		2.81	.02	0.4	4.0	
Separation = 19.09	Reliability = 1.00					

Misfitting criteria: MnSq > 1.4 and ZStd > 2.0
I = Instrumental ADL P = Personal ADL

Results of the hierarchical linear continuum presented in Figure 2 indicated that the IADL items were harder for the Alzheimer patients than the PADL items. As can be seen in Figure 2, the DAQ items are spread out over a continuum of ability ranging from 4.91 to -4.04 logits with finances the hardest item and eating the easiest. Some items were of similar difficulty for our Alzheimer's subjects (e.g., shopping & cooking, dressing & bathing). A high item separation, 19.09 with a corresponding reliability = 1.0, indicates that items of the DAQ were separated into almost 20 different difficulty levels by the persons being tested. A high person separation, 3.85 with a corresponding reliability = 0.94, indicates that these patients could be separated into 4 different independent levels by the DAQ. Combining these two separations, the DAQ worked well to sensitively measure the independence in ADL of the sample population.

Figure 2 also shows that the items are well targeted to the subjects' level of ADL independence. Subjects located toward the top of the scale are more independent in ADL than those located toward the bottom. It can be seen that most subjects fall within the range of the item difficulty. Thirty of the 276 subjects appear to have more than 50% of probability to require assistance with the easiest item—eating.

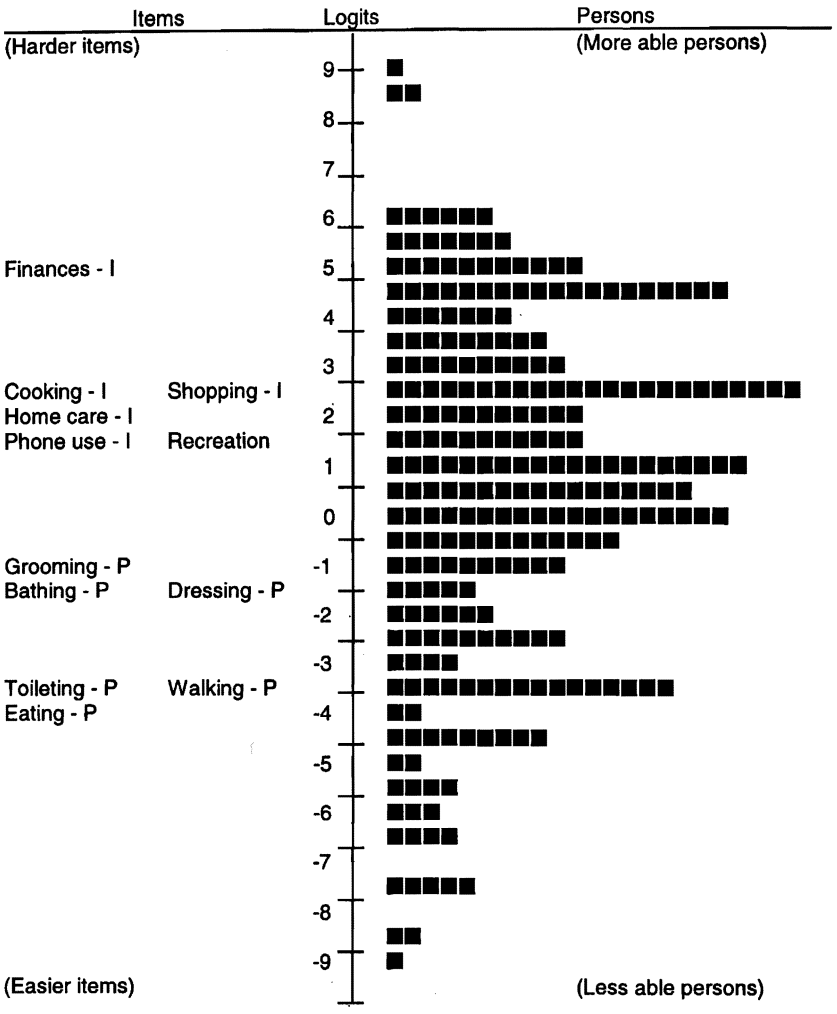


Figure 2. Distribution of Alzheimer's subjects (N = 276) and DAQ items. Person separation is 3.85 (reliability = 0.94); item separation is 19.09 (reliability = 1.0)

Discussion

Typically, examiners use ADL assessments without ensuring the quality of the rating scale structure or if the assessment items meet the

criteria of unidimensionality. The purpose of our study was to evaluate the internal validity, construct validity, and rating scale structure of the DAQ, a measure devised to assess ADL independence in people with Alzheimer's disease. We examined the rating scale structure and unidimensionality of the DAQ using a Rasch Measurement approach. This study provides the evidence that a 5-point rating scale was more appropriate than our original 10-point rating scale, underscoring the importance of evaluating the underlying rating scale structure. When we restructured the DAQ to a 5-point scale and excluded two items which were found to be misfits (see below), the remaining items were positioned along a hierarchical continuum, with IADL tasks more difficult than PADL tasks. This ordering is consistent with known developmental milestones and the reversal of that developmental hierarchy during the clinical progression of Alzheimer's disease.

When we examined if the 14 DAQ items worked together to define a single construct of ADL independence, we found that all but two items (sleeping and walking) measured the same underlying construct and thus could be combined to generate a single measure of ADL independence. Because "sleeping" and "walking" had fit statistics outside the suggested range (i.e., misfits), we concluded that sleeping did not define the same construct as the rest of the items while "walking" needs to be reworded. It may be that "sleeping" is more a reflection of an underlying biological function than an ADL. We consider "walking" a crucial item for the DAQ. The fact that it misfit suggests that the item might need to be more clearly defined. We hypothesize that the ambiguity of its definition resulted in rater confusion. The term "ambulating" might be more appropriate than "walking" in a functional sense. An option would be to redefine "walking," re-administer the questionnaire, and reexamine the data to see if the revised item demonstrated better fit to the measurement model. The current study suggests that IADL and PADL together define a single construct, ADL, with IADL being more challenging than PADL (Shown in Figure 2). This result lends credence to Doble and Fisher's (1998) argument against the usual practice of generating separate PADL and IADL measures for subjects.

Why is this finding of unidimensionality with the DAQ important? Assessment measures with sound psychometric properties such as the DAQ can provide quantitative data for research studies or a clinical tool for planning treatment appropriate to the patient. Knowledge of the task difficulty and scale structure can provide a sequential guide for planning

intervention programs. For instance, if Mr. Smith has an ADL independence measure of -1.0 on the DAQ (Figure 2 and Table 2), we would expect him to have about 50% of probability to demonstrate independence in grooming (Measure = -1.17, Error = 0.16), have more than 50% of probability to demonstrate independence in dressing, (Measure = 1.49, Error = 0.15) and have almost no problem in demonstrating independence in eating (Measure = -4.04, Error = 0.15). The treatment goal would be to maximize Mr. Smith's potential for independence in grooming and maintain his independence in other items with measures smaller than grooming. Such ADL profiling is useful for clinical management strategies and has been employed in this manner at the NIH for some time. In addition, given the progressive nature of Alzheimer's disease, we have found the DAQ useful for tracking changes in ADL independence in longitudinal studies of this population.

In summary, the results of this study support that the modified DAQ is a valid scale and a viable measure of ADL independence in people with Alzheimer's disease. This information allows us to have greater confidence in using it to gather data for making initial and on-going assessment of patients' ADL status, for guiding decisions about therapeutic interventions and for measuring outcomes of our interventions in both clinical situations and research studies. Such use of the DAQ is currently ongoing.

References

American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders*. Washington, D.C.: American Psychiatric Association.

Doble, S. and Fisher, A.G. (1998). The dimensionality and validity of the older americans resources and services (OARS) activities of daily living scale. *Journal of Outcome Measurement, 2*, 4-24.

Linacre, J. M. (1998). Visual analog scales. *Rasch Measurement Transactions, 12*, 639.

Linacre, J. M. (1996). *FACETS: for PC-Compatibles*. Chicago: MESA Press.

Linacre, J. M. (1998). *A User's Guide to FACETS*. Chicago: MESA Press.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D. and Stadlan, E. (1984). Clinical diagnosis of Alzheimer's disease. Report of the NINCDS-ADRDA work group. *Neurology, 34*, 939-944.

Nunnally, J. C. (1967). *Psychometric Theory*. New York: McGraw Hill.

Oakley, F., Sunderland, T., Hill, J., Phillips, S., Makahon, R., and Ebner, J. (1991). The daily activities questionnaire: A functional assessment for people with Alzheimer's disease. *Physical and Occupational Therapy in Geriatrics, 10,* 67-81.

Wright, B. D. and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8,* 370.

Wright, B. D. and Masters, G. N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

JOURNAL OF OUTCOME MEASUREMENT®, *3*(4), 308–322
Copyright© 1999, Rehabilitation Foundation, Inc.

Mapping Variables

Mark H. Stone
Adler School of Professional Psychology

Benjamin D. Wright
The University of Chicago

A. Jackson Stenner
MetaMetrics, Durham, NC

This paper describes Mapping Variables, the principal technique for planning and constructing a test or rating instrument. A variable map is also useful for interpreting results. Modest reference is made to the history of mapping leading to its importance in psychometrics. Several maps are given to show the importance and value of mapping a variable by person and item data. The need for a critical appraisal of maps is also stressed.

Requests for reprints should be sent to Mark H. Stone, Adler School of Professional Psychology, 65 East Wacker Place, Suite 2100, Chicago, IL 60601-7203, e-mail: MHS@Adler.edu.

The **Map of a Variable** is the beginning and end of assessment. But we must immediately add that variable construction never ends because it is never complete; it is ever continuing. Variables require continuous attention for their development and maintenance. The map of a variable is a visual representation of the current status of variable construction. It is a pictorial representation of the "state of the art" in constructing a variable.

The Origins of Mapping

Maps are visual guides. They ground us in a stable frame of reference and give a sense of direction. How frequently we use expressions of belief implying vision, "Do you see?", "Now I see.", "Show me what you mean.", and "Put me in the picture." These expressions testify to the visual power inherent in pictorial representations and conveyed in speech and writing. Mapping visualizes the extent of our knowledge.

Maps are indispensable to planning and traveling. Map making has great utility. The inability to understand or make use of maps is a handicap to understanding the world.

The earliest maps used naturally occurring phenomena—celestial and terrestrial—to identify features. If we look at the sky on a starry night we can use the "pointers" of the Big Dipper to locate Polaris, the pole star. Although both dippers move, they rotate around Polaris which appears fixed and determines north. From this "fixed" star we orient ourselves to the points of the compass. More comprehensive maps of the heavens include the popular constellations of the Zodiac, lesser known constellations and other celestial features. The more celestial features we know, the better oriented we become to a starry night.

Terrestrial markers also serve to orient. A lake, a river, or a mountain may be used to anchor locations. Celestial and terrestrial maps have been used for centuries and are sometimes brought into relationship with one another. Today's roadmaps are but a current update of the state of knowledge in local geography. A map is an analogy, an idea that pictures an abstraction. While the map may initially seem superficial, incomplete, or even inaccurate, it still serves a purpose. The map shows the current status of what is known about a domain.

Maps by their very nature, invite improvement. Every edition of a map calls attention to its accuracy and inaccuracy. Each new edition incorporates changes from a previous one resulting in a new and more accurate version.

Consider a map with the lines of longitude and latitude. This illustrates the benefits of superimposing an abstraction upon the natural contours of land and sea. Abstractions enhance maps by expediting generalization.

Natural reference points also explain by serving as markers to ground our observations. The more natural reference points we can employ, the fewer the resulting errors. The wider apart the natural markers, the greater the possibility of error. Lloyd Brown (1949) provides a comprehensive history of map building with numerous illustrations that record how maps have become increasingly more accurate. John Wilford (1981) has produced a similar, but more recent history. Edward Tufte's recent publications (1979, 1983) offer a panorama of useful visual strategies together with his critique of how visual displays can facilitate the interpretation of data or mislead.

The use of maps illustrates several important aspects:

1. Maps are useful pictures of experience.
2. Inaccuracies are successively and inevitably corrected.
3. Abstractions, such as longitude and latitude, enhance mapping.
4. More knowledge produces greater accuracy.

Maps of Variables

Map topography is a useful application to psychometrics because a map is an abstraction of a variable. The variable implied by a test can first be pictured as a line (Wright and Stone, 1979, pp. 1-6). It is a line with direction illustrated by an arrow. The variable is defined by items and persons, but other useful characteristics can also be incorporated on the map. Continuous improvement is irresistible. Maps invite further corrections. The more information we gather about the variable, the more accurate our representation becomes. Finally, this pictorial representation of the variable invites yet further abstractions that generalize understanding.

Rudolph Carnap wrote,

> The nineteenth-century model was not a model in this abstract sense (i.e., a mathematical model). It was intended to be a spatial model of a structure, in the same way that a model ship or airplane represents an actual ship or plane. Of course, the chemist does not think that molecules are made up of little colored balls held together by wires; there are many features

> of his model that are not to be taken literally. But, in general spatial configuration, it is regarded as a correct picture of the spatial configuration of the atoms of the actual molecule. As has been shown, there are good reasons sometimes for taking such a model literally—a model of the solar system, for example, or of a crystal or molecule. Even when there are no grounds for such an interpretation, visual models can be extremely useful. The mind works intuitively, and it is often helpful for a scientist to think with the aid of visual pictures. At the same time, there must always be an awareness of a model's limitations. The building of a neat visual model is no guarantee of a theory's soundness, nor is the lack of a visual model an adequate reason to reject a theory. (*Carnap, 1966, p. 176*)

Carnap's exposition clearly indicates the value of a map in fostering pictures by which to visually conceptualize an intuitive idea. He also cautions that maps are not substitutes for reality, but pictures and as such they cannot be interpreted literally.

Using Maps

There are three uses of maps:

To DIRECT... where we are planning to go,

to LOCATE... where we are, along the way, and

to RECORD... where we have been.

These three uses indicate that a map is the beginning and end of test construction. In the beginning stages, a map defines our intentions. At the end, it is a realization of progress to date. In between are markers along the way. Maps of variables are never finished because they invite constant correction and improvement. When maps embody abstractions derived from experience they connect the world of the mind to the world of experience. Abstraction is validated by correspondence to experience and experience is understood by abstraction.

Mapping illustrates the dialogue that must take place between these two worlds in order to communicate constructively. A map is a visual, operational definition of a variable. While maps are necessarily only models, their pictorial representation invites continual correction, ever increasing their accuracy.

Graphs as Maps

Graphs of functions are maps showing the relationship between two variables. Graphs make it easy to see by looking whether a useful function is emerging.

Graphs make functions recognizable and familiar. We recognize linearity in a straight line, and special curves describe a parabola or a quadratic relation. Other well- known functions like the undulating curve for the sine are easily recognized by their shape. The graphs of functions are maps as familiar to their users as roadmaps are to motorists. They aid understanding by simplifying the process and allowing us to "see" a complex representation.

A Map is an Analogy

Measurement is made by analogy. Our most efficient and utilitarian measures rely upon visual representation. The ruler, the watchface, the mercury column, and the dial are common analogies used to record length, time, temperature, and weight. The utilitarian success of analogy in these measuring tools is demonstrable by their ubiquity.

1. The "intended map" of the variable is the idea, plan, and best formulation of our intentions.
2. The "realized map" of the variable which is made from item calibrations and person measures implements the plan.
3. Continuous dialogue between intention (idea) and realization (data) produces and maintains the validity of the variable.
4. A "Map of a Variable" is the scope and sequence of instruction because it shows how to sequence instruction and how to relate it to assessment.
5. Progress from instruction and resultant learning can be located on the variable. Growth can be seen and measured.
6. There are shortcomings to maps, especially evident if their use is "pushed" to extremes.

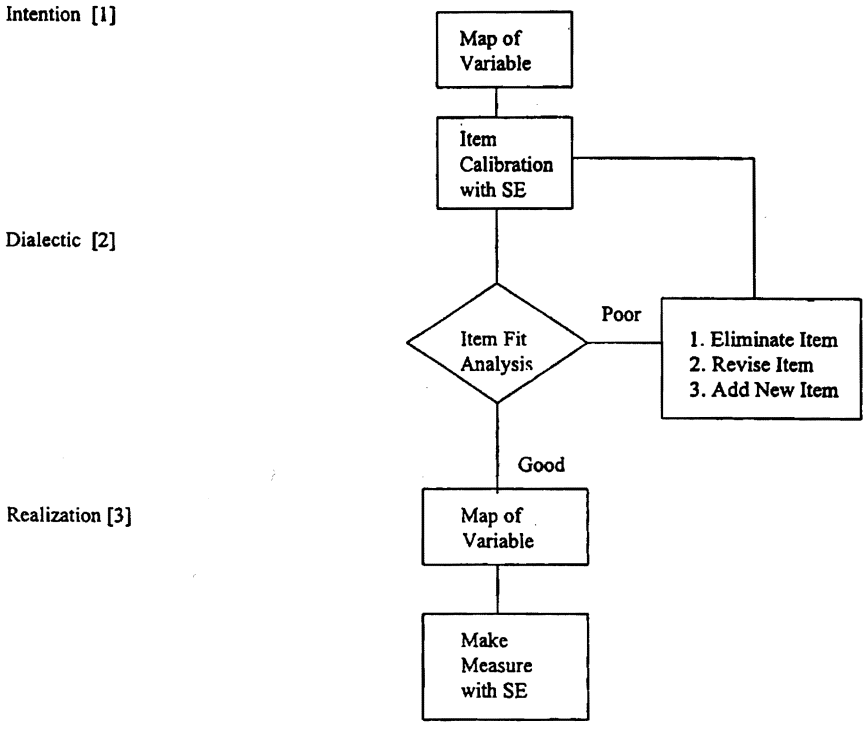
What writers like Carnap (1966) and Kaplan (1964) present in their discussions about the shortcomings of models also applies to maps. We must be careful not to expect too much of a map and ascribe more substance to what is produced than can be justified. Constant monitoring of map building is necessary. Monmonier's (1996) book "How To Lie With Maps" presents in a useful and amusing way the fallacies that can result from viewing a map as a "finished product" rather than as a "fiction," an

approximation of the outcome and one that is in process and never completed. Braithwaite (1956) has also cautioned, "The price of the employment of models [maps] is eternal vigilance."

Psychometric maps serve as the plan for instrument development and revision. The map of a variable is a blueprint for a test. When a map is logical and well constructed, its implementation can be straightforward in the form of ordered items.

Figure 1 is a flowchart of the steps in bringing a variable into existence. Its development is guided by a map of intention.

The Stages of Mapping a Variable



1. The "intended map" of the variable is the idea and plan of our intention.
2. There is continuous dialogue between the intention (idea) and realization (data) to maintain validity and quality control. The degree of correspondence between the maps of intention and realization indicates the degree of success achieved.
3. The "realized map" of the variable conveys in the item calibrations and person measures the best outcome to date.

Figure 1. Flowchart for Variable Construction

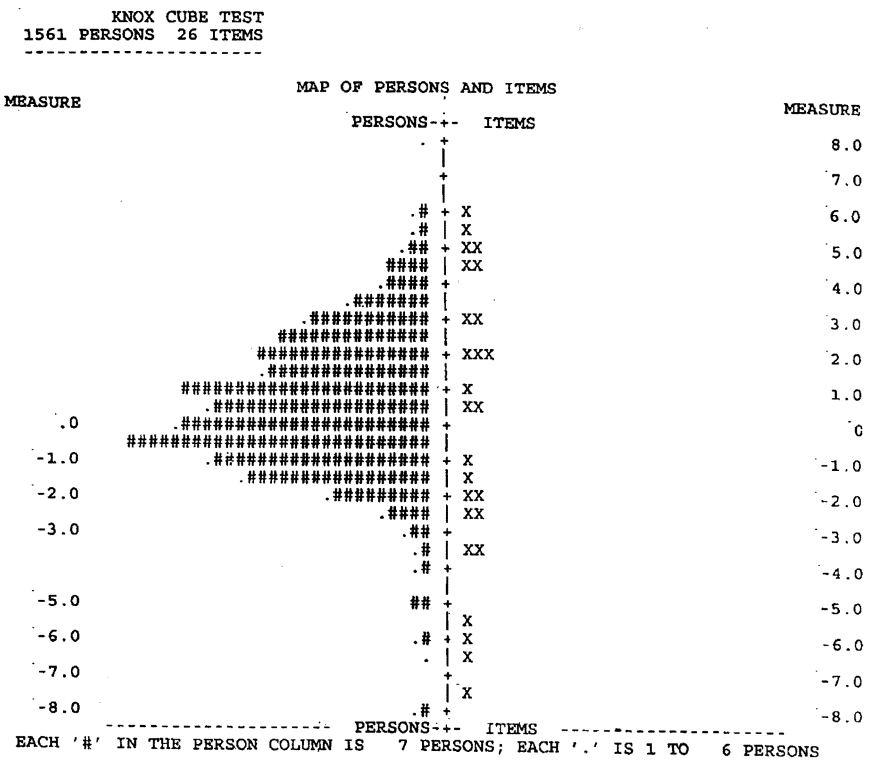


Figure 2. Item/Person Map. From Stone and Wright (1980) Knox Cube Test. Itasca, IL: Stoelting.

When the map is empirically verified, it documents a successful realization of an idea. The map of the variable pictures both the idea and its realization in the form of calibrated items and measured persons (Wright and Stone, 1996, Chapter 14). It embodies the construct validity of the instrument.

Figure 2 is the item/person map for the Knox Cube Test (Stone and Wright, 1980) generated by BIGSTEPS (Wright and Linacre, 1991 to date) This map as well as others generated from WINSTEPS (Linacre, 1999) greatly assist the psychometrist in variable construction. However, it is simple maps like this one that make psychometric analysis understandable to content specialists and other persons interested in the results, but not concerned about methodology.

Binet's work in test development began more than 100 years ago. His work implies mapping although he did not employ the term.

> First of all, it will be noticed that our tests are well arranged in a real order of increasing difficulty. It is as the result of many trials, that we have established this order; we have by no means imagined that which we present. If we had left the field clear to our conjectures, we should certainly not have admitted that it required the space of time comprised between four and seven years, for a child to learn to repeat 5 figures in place of 3. Likewise we should never have believed that it is only at ten years that the majority of children are able to repeat the names of the months in correct order without forgetting any; or that it is only at ten years that a child recognizes all the pieces of our money. (*Binet, 1916, p. 185*)

Binet clearly indicates how data from experience was used to established a hierarchy of item difficulty. He makes special note of the requirement for "well-arranged" items expressing a "real order". Binet also relied on "numerous" replications of ordered items in order to produce the level of accuracy he desired.

> One might almost say, 'It matters very little what the tests are so long as they are numerous'. (*Binet, 1916, p. 329*)

Binet clearly stressed (1) item arrangement by difficulty order, (2) numerous items, sufficient for precision. How else can one be successful? There is no other way except to do as Binet did: begin with an idea for a variable, illustrate the variable by items, arrange them by their intended difficulty, and measure persons by their locations among the items. The hallmark of Binet's efforts is his early effort at benchmarking items and persons on a variable. He must have had a mental map of what he intended, although there is no indicate of one in his writings.

An early example of a psychometric map is Thurstone's "Scale of Seriousness of Offense" (1927, 1959) shown in Figure 3. His map marks out the severity of offenses from "vagrancy" located at the bottom end to "rape" at the top.

The scale is further subdivided into three offense categories: (1) sex offenses, located at the top of the scale, (2) injury to the person, located from the top to the middle, and (3) property offenses, located from the middle of the scale and down. Thurstone's map provides insight into a hierarchy of criminal acts and a practical "ruler" for determining, not only the location of offenses, but the "distance" between them.

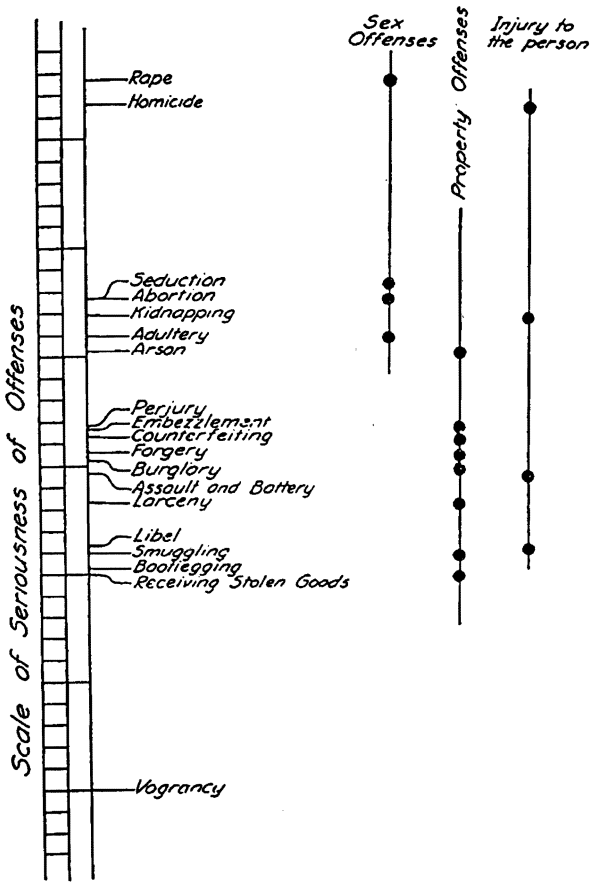


Figure 3. Scale of Seriousness of Offense. From L.L. Thurstone (1959). The measurement of values. Chicago: The University of Chicago Press, p. 75.

Figure 4 is a map of an achievement variable: WRAT3 (Wilkinson, 1993). This test of achievement measures (1) word naming, (2) arithmetic computation, and (3) spelling from dictation. Items are arranged according to difficulty. These maps progress from left to right indicating increases in item difficulty and person ability. The arrangement of items indicates the expected arrangement of persons according to their abilities. Less able persons will be located to the left of more able persons. Able persons will find the items on the left easier than those items further along to the right. These three variables follow developmental lines of learning, correspond to instructional goals and make test administration efficient and informative. The map of each WRAT3 variable is enhanced by sample items illustrating progressive diffi-

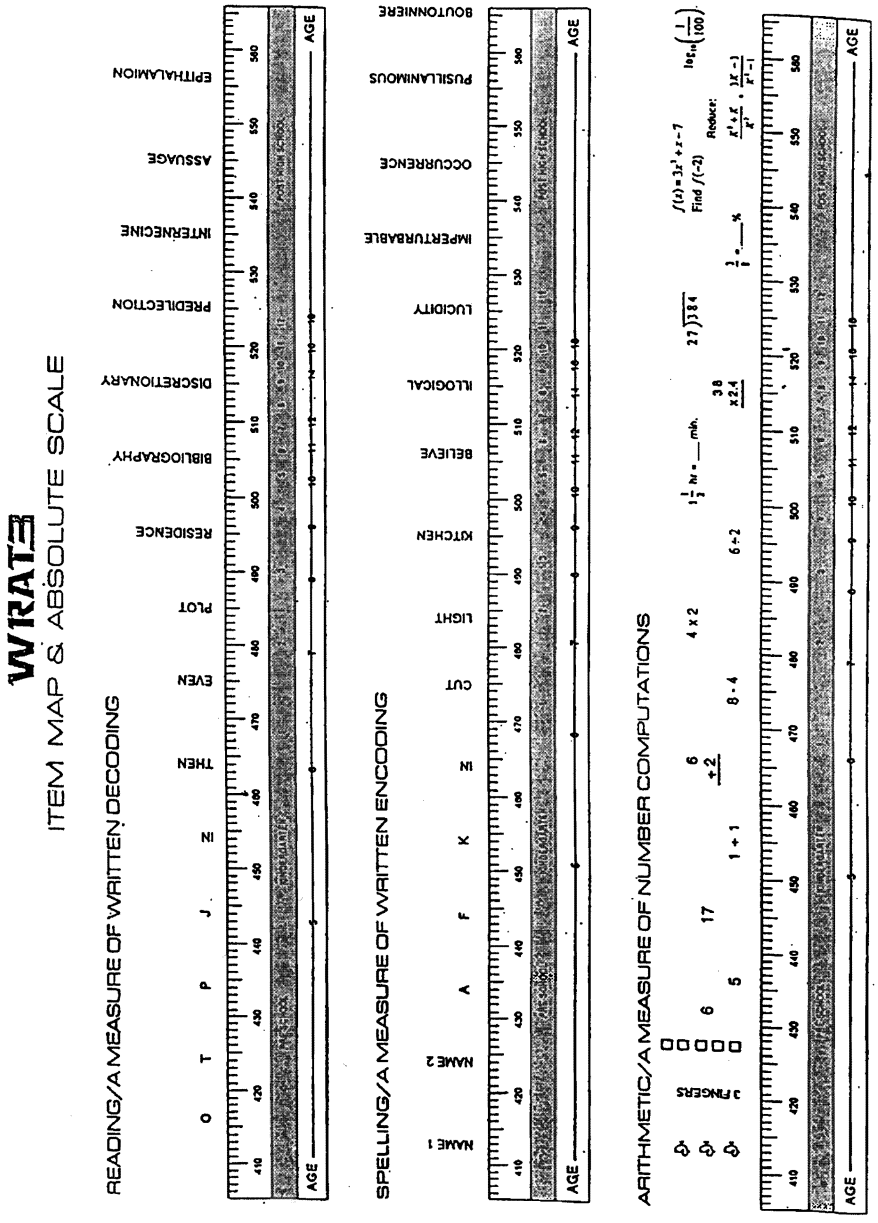


Figure 4. Wide Range Achievement Test: WRAT3. From G. S. Wilkinson (1993). The Wide Range Achievement Test (1993 Edition). Wilmington, DE: Wide Range.

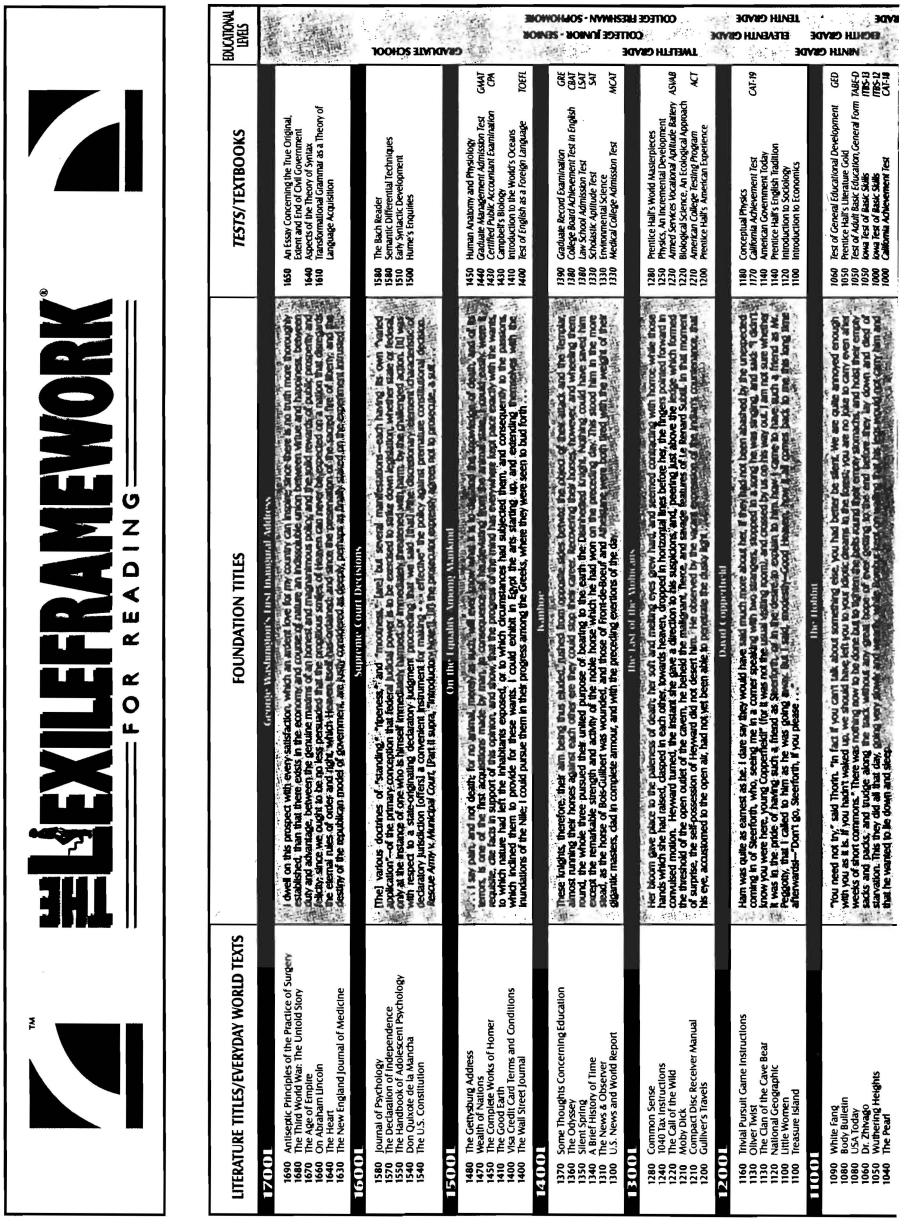


Figure 5. The Lexile Framework. From A. J. Stenner. (1995). Lexile Research Triangle Park, NC.

culty and below the items is an equal-interval scale indicating the measures. The locations of successive average grade levels are given as well as the average age associated with the item calibrations and person measures. This map is a succinct picture of the three WRAT3 variables.

These maps have immediate application. Like the marks made on a door jamb to show the increasing height of a child, this map shows student progress on three achievement variables. The maps show order to the items and measures. Progress of pupils along this educational ruler is enhanced by criterion and normative locations. The grade and age norms show growth. The map provides useful information to students, teachers and parents.

Figure 5 is a reduction of the "map" of the Lexile Scale of Reading© copyright 1994, Metametrics).

The master map is larger and more comprehensive and requires a chart greater than 2' by 3' in order to picture only some of the large amount of available information. Lexile calibration values have been computed for a substantial number of trade books, texts, and tests. The title column indicates the content validity of the scaling. The educational levels column shows the increase in difficulty corresponding to reading more difficult materials. Construct validity can be demonstrated by these map locations. Educational levels, ages, and other information can be positioned on the Lexile Map. Criterion and construct validity are demonstrated by these relationships.

Mapping technology offers a powerful tool for conjointly ordering objects of measurement i.e. readers and indicants i.e. texts. Meaning accrues as this conjoint ordering of reader and text is juxtaposed with other orderings including grade level, income or job classification. Collections of these "orderings" constitute a rich interpretive framework for bringing meaning to the measurement of human behavior.

A good leap in understanding and utility is accomplished when the ordering of indicants along the line of the variable can be predicted from theory. In every application of physical science measurement, instrument calibration is accomplished via theory not data. Social science measurement stands alone in its reliance on data in the construction of instrument calibrations and co-calibrations between instruments.

Perhaps the key advantage of theory based calibrations is that an absolute framework for measure interpretation can be constructed without reference to any individual or group measures on objects or indicants. The

prospect of absolute measurement, long taken for granted in the physical sciences, has until recently eluded social scientists. The building of maps for the major dimension of human behavior is now possible because of the theoretical work of Rasch and Wright, amplified by the work of colleagues.

One pretender to the kind of mapping process outlined above is evident in NAEP's use of the Reading Proficiency Scale (RPS). The RPS is a transformed Rasch scale with an operating range of 0 to 500. NAEP describes performance at grades 4, 8, and 12 as rudimentary, basic, intermediate, adept or advance depending upon the RPS attained by each student. Thus, a rudimentary reader has an RPS = 150, an intermediate reader at RPS = 250 and an advanced reader at RPS = 350. So far so good, since all we have done is "name" certain "anchor" points on the RPS scale.

Problems develop when reader performance on the RPS scale is described using relative language such as stating that a rudimentary reader "can follow brief written directions" or "can carry out simple, discrete reading tasks" or a basic reader "can understand specific or sequentially related information." An intermediate reader "can search for specific information, interrelate ideas, and make generalizations." An adept reader "can analyze and integrate less familiar material and provide reactions to and explanations of the text as a whole." An advanced reader "can understand the links between ideas even when those links are not explicitly stated."

These statements are not appropriate descriptions of scale points along the RPS scale. Rather they are good descriptions of the behavioral consequences of more or less accurately matching the demands of a text with the capabilities of a reader. Thus, rather than describing absolute scale positions, these annotations, in fact, describe differences between a reader measure and a text measure. When a text's measure exceeds a reader's measure, comprehension is low and the kinds of reader behaviors used above describe a "basic" result.

When a reader's measure exceeds a test's measure the kinds of reader behaviors used to describe adept and advanced readers are evident. The key point is that each of these behaviors can be elicited in the same reader simply by altering the level of text that is presented to the reader. Thus we can make a 400L (second grade level) reader adept by presenting a 100L text or a 400L reader rudimentary by presenting 800L text. Comprehension rate is always relative to the match between reader and text and it is this rate, rather than the reader's measure, that is appropriately

described in behavioral and proficiency terms. Much confusion has resulted from a failure to recognize this distinction.

Summary

Successful item calibration and person measurement produces a map of the variable. The resulting map is no less a ruler than the ones constructed to measure length.

The map indicates the extent of content, criterion, and construct validity for the variable. The empirical calibration of items and the measures of persons should correspond to the original intent of item and person placement, but changes must be made when correspondence is not achieved. There should be continuous dialogue between the plan, person measures, and item calibrations. Variables are never created once and for all. Continuous monitoring of the variable is required in order to keep the map coherent and up-to-date. Support for reliability and validity does not rest in coefficients, but in substantiating demonstrations of relevant and stable indices for items and measures. Such indications must be continuously monitored in order to maintain the variable map and assure its relevancy.

References

Binet, A. (1916). *The development of intelligence in children.* Baltimore: Wilkins and Wilkins.

Braithwaite, R. (1956). *Scientific explanation.* Cambridge: Cambridge University Press.

Brown, L. (1949). *The story of maps.* New York: Little, Brown and Company.

Carnap, R. (1966). *Philosophical foundations of physics: An introduction to the philosophy of science.* New York: Basic Books.

Kaplan, A. (1964). *The conduct of inquiry.* New York: Thomas Crowell.

Linacre, J. (1999). *WINSTEPS.* Chicago: MESA Press.

Metametrics. (1995). *The lexile framework.* Research Triangle Park, NC: Author.

Monmonier, Mark. (1996). *How to lie with maps.* (2nd ed.). Chicago: University of Chicago Press.

Stone, M. (1995). Mapping variables. Paper given at the Midwest Objective Measurement Seminar. Chicago: The University of Chicago.

Stone, M. and Wright, B. (1980). *Knox's Cube Test.* Itasca, IL: Stoelting.

Thurstone, L. (1959). *The measurement of values.* Chicago: The University of Chicago Press.

Tufte, E. (1983). *Visual explanations: Images and quantities, evidence and narrative*. Chesire, CT: Graphics Press.

Tufte, E. (1997). *The visual display of quantitative information*. Chesire, CT: Graphics Press

Wilkinson, G. (1993). *Administration Manual: WRAT-R*. Wilmington, DE: Wide Range, Inc.

Wilford, J. (1981). *The mapmakers*. New York: Alfred Knopf, Inc.

Wright, B. and Linacre, J. (1991). *BIGSTEPS*. Chicago: MESA Press.

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

Wright, B. and Stone, M. (1996). *Measurement essentials*. Wilmington, DE: Wide Range, Inc.

JOURNAL OF OUTCOME MEASUREMENT®, *3*(4), 323–338
Copyright© 1999, Rehabilitation Foundation, Inc.

Many-facet Rasch Analysis with Crossed, Nested, and Mixed Designs

Randall E. Schumacker
University of North Texas

Many-facet Rasch analysis provides the bases for making fair and meaningful decisions from individual ratings by judges on tasks. The typical measurement design employed in a many-facet Rasch analysis has judges crossed with other facets or conditions of measurement. A nested design does not permit facets to be compared. However, a mixed design can be used to achieve a common vertical ruler when the frame of reference permits commensurate measures to be linked. Examples of crossed, nested, and mixed designs are presented to illustrate how a many-facet Rasch analysis can be modified to meet the connectivity requirement for comparing facet measures.

Requests for reprints should be sent to Randall E. Schumacker, Department of Technology and Cognition, Matthews Hall 304, University of North Texas, Denton, Texas 76203-1337, e-mail: rschumacker@unt.edu.

The many-facet Rasch model implies that person measures are obtained under measurement conditions involving different judges and tasks. A complete set of measures involving all judges rating all persons on one or more tasks can be obtained, or different judges can rate different persons on one or more tasks. A many-facet design involving fewer ratings by different judges on a set of tasks however must meet certain connectivity requirements (Linacre, 1994a). When measures are connected, a common vertical ruler is created which permits facet comparisons; otherwise separate vertical rulers are created for each facet or measurement condition.

The many-facet Rasch analysis computes fair person measures from ratings by judges (Linacre, 1994a). The basic many-facet Rasch model is: $\text{Log}\,(P_{nijk} / P_{nij(k-1)}) = B_n - D_i - C_j - F_k$, where:

P_{nijk} = probability of student n on task i by rater j being given a rating of k

$P_{nij(k-1)}$ = probability of student n on task i by rater j being given a rating of k - 1

B_n = ability of student n

D_i = difficulty of task i

C_j = severity of rater j

F_k = difficulty of threshold across rating scale categories from k-1 to k.

Each facet (i.e., judge or task) is assumed to be independent from the other facets. The facets combine to give the probability for a person's rating by a particular judge. In order for the facets to be compared, they must be on the same linear scale. This requires the creation of a vertical ruler. If a common vertical ruler can be created, then a "subset connection ok" message appears. However, if this is not possible then separate vertical rulers are created, one for each facet. A **completely** nested design would yield separate vertical rulers, one for each facet, because the facets can not be connected. To illustrate, a crossed design, a nested design, and a mixed design are presented.

Design Comparisons

Crossed Design

A crossed design would have facet elements as depicted in Table 1. There are three judges who rated three students on three tasks (A, B, and C).

Table 1

Crossed Design

	Judge		
	1	2	3
Task	A B C	A B C	A B C
Student			
1	3 4 5	4 2 5	4 3 5
2	2 5 4	4 3 5	4 4 3
3	3 4 5	2 3 4	5 4 3

Nested Design

A nested design would have facet elements as depicted in Table 2. There are three judges, but each judge only rates the students on one task. Judge 1 only rates task A, judge 2 only rates task B, and judge 3 only rates task C. The basic requirement of connectivity (linking) in many-facet analysis does not exist because the ratings are nested within each task.

Table 2

Nested Design

	Judge		
	1	2	3
Task	A B C	A B C	A B C
Student			
1	3 - -	- 2 -	- - 5
2	2 - -	- 3 -	- - 3
3	3 - -	- 3 -	- - 3

Mixed Design

A way to achieve connectivity for the creation of a common vertical ruler is to have at least one judge crossed with all elements of a facet, i.e., task. For example, a fourth judge could be added to the nested design data yielding a mixed design (Table 3). The fourth judge, however, would have to rate all students on all three tasks.

Table 3

Mixed Design

	Judge											
	1			2			3			4		
Task	A	B	C	A	B	C	A	B	C	A	B	C
Student												
1	3	-	-	-	2	-	-	-	5	2	4	5
2	2	-	-	-	3	-	-	-	3	2	3	5
3	3	-	-	-	3	-	-	-	3	3	4	4

Program and Data Comparisons

Two programs are individually executed to input and output the various files needed for a many-facet Rasch analysis. The first program, **FACFORM** (Linacre, 1994b), is executed to input a many-facet program (*.key) which reads the raw data file (*.asc) and creates a facet specification program file (*.spe) and a facet data file (*.fac). The raw data file (*.asc) depicts how the data needs to be formatted for the crossed, nested, and mixed designs. A judge number appears as the first line of data in the raw data file (*.asc). The next three lines of data are for the students; coded as 1, 2, or 3. After each student, a task number and rating is entered for the three tasks. Since numeric data is expected, Task A = 1, Task B = 2, and Task C = 3. The second program, **FACETS** (Linacre, 1994b), is executed to input the facet specification program (*.spe) and read the facet data file (*.fac). The programs and data files for the crossed, nested, and mixed designs are listed in the Appendix. The results are listed in a facet output file (*.out).

Result Comparisons

Many-facet Rasch analysis output from the facet output files (*.out) for the crossed, nested, and mixed designs is presented in an abridged format. The chi-square statistics are not reported since connectedness (linkage) of the facets is not a problem related to model fit or facet significance. See Schumacker and Lunz (1997) for a discussion and interpretation of the different chi-square statistics output in many-facet Rasch analyses.

Crossed Design

An important message that appears when the facets are connected (linked) is: "Subset connection OK". A table (Table 4) is produced that indicates the common vertical ruler.

Table 4

All Facet Vertical Rulers

Measure	Judge	Student	Task	Scale
1				(5)
				4
	Judge1 Judge3		A	
			B	
		Student3		
	Judge2			
0		Student2		-----
		Student1		
				3
			C	
-1				(2)

Nested Design

When analyzing the data in the nested design, the following message appears: "Warning! 3 disjoint subsets are group-anchored". This message indicates that the three tasks are "disjointed" or not connected and "anchored" to their unique element value for the facet. You can bypass this warning by including the command statement "Subsets=Bypass" in the many-facet program, but it is **not** recommended when data are nested.

Three separate vertical ruler tables are listed (Tables 5, 6, and 7), one for each judge's ratings on a unique task, as follows:

Table 5

Disjoint Subset 1 Vertical Summary

Measure	Judge	Task	Scale
3			(5)
	Judge3		
2			
1			
0		C	3
-1			
-2			
-3			(2)

Table 6

Disjoint Subset 2 Vertical Summary

Measure	Judge	Task	Scale
3			(5)
2			
1			
0		B	3
-1			
-2			
	Judge2		
-3			(2)

Table 7

Disjoint Subset 3 Vertical Summary

Measure	Judge	Task	Scale
3			(5)
2			
1			
0		A	3
-1			
-2			
	Judge1		
-3			(2)

Mixed Designs

Adding a judge who rated all students on all tasks created the mixed design. The additional ratings by this fourth judge permitted the facets to be compared on a common vertical ruler. The connectivity requirement was met by having two ratings on each task by two judges. The program output indicated: "Subset connection O.K.", but gave the message, "Warning! Estimates may have not converged". The addition of more ratings by judges on all tasks (essentially more data) would remove the message. Table 8 presents the common vertical ruler for the mixed design analysis.

Table 8

All Facet Vertical Rulers

Measure	Judge	Student	Task	Scale
5			A	(5)
4				
3				
2	Judge1			4
1		Student2		
0	Judge4			
		Student3		
		Student1		
-1			B	3
-2	Judge3			
-3			C	
-4	Judge2			
-5				(2)

Conclusions

The crossed design example contained three judges who rated three students on three tasks. The nested design example contained three judges, but each judge only rated the three students on a single task. The mixed design added a fourth judge to the nested design with the requirement that all students were rated on all tasks. This provided the minimum linking requirement of two ratings per task by the judges. Obviously, more than one judge who rated all students on all tasks could have been included. More complex nested designs could be modified as long as the linkage requirement is met. Whether facets are intended to be linked, that is, **not** remain nested in a design, is a theoretical issue, beyond the scope of this article.

Many-facet analyses contain measurement conditions that are crossed. However, **complete** data in a crossed design is not necessarily required (see Linacre, 1994a). Data in a nested design doesn't have the required "linking" characteristics to permit the creation of a common vertical ruler for comparing facet measures, so only interpretation of an individual judge's ratings on a single task is possible. A mixed design is a modification of the nested design such that at least one judge is crossed with all elements of a facet, thus permitting linkage (connectivity). Linkage is a basic requirement in producing the common vertical ruler in many-facet analyses. If the measures are commensurate with each other, then the mixed design technique can be used to link the facets for comparative purposes.

References

Linacre, J. M. (1994a). *Many-Facet Rasch Measurement.* MESA Press: Chicago, Illinois.

Linacre, J. M. (1994b). *A User's Guide to Facets: Rasch Measurement Computer Program.* MESA Press: Chicago, Illinois.

Schumacker, R. E. and Lunz, M. E. (1997). Interpreting the chi-square statistics reported in the many-faceted Rasch model. *Journal of Outcome Measurement, 1*(3), 239-257.

Appendix

Crossed Design

CROSSED.KEY

```
; File: crossed.key
; This program creates a facet specification file and
comma separated data file

$Input    = crossed.asc     ; ascii data file
$Output   = crossed.fac     ; comma separated data file
$Spoutput = crossed.spe     ; facet specifications file

$Facets = 3                 ; judge, student, and task
(crossed design)

; labels for the Facets

$Flabel = 1, "judge"              ;  3 judges
$Flabel = 2, "student"            ;  3 students
$Flabel = 3, "task"               ;  3 tasks

; FACETS specifications

$Spec = "Title = Crossed Design"
$Spec = "Output = crossed.out "   ;the FACETS output file
$Spec = "Models = ?,?,?,R5"       ; 3 facets using 5 point
rating scale

; read judge number

check1 = $C1
judge=1

$DO = (check1 <>"") ; perform $DO while judges exist
$Label = 1,judge    ; first comma separated value

; skip to next line
$Nextline

; read student number
check2 = $C2
student=1

$DO = (check2 <>""); perform $DO while students exist
$Label = 2,student ; first comma separated value on next
line
```

(Appendix continued on next page)

(Appendix continued from previous page)

```
; read in three tasks and associated ratings
  $Label=3,$S4W1       ; Task A - col 4
  $Rating = $S7W1      ; rating - col 7
  $Label=3,$S9W1       ; Task B - col 9
  $Rating = $S12W1     ; rating - col 12
  $Label=3,$S14W1      ; Task C - col 14
  $Rating = $S17W1     ; rating - col 17
$Nextline
  check2=$C2
  student=student+1
$Again
  check1 = $C1
  judge=judge + 1
$Again
```

CROSSED.ASC

```
1,
1, 1, 3 2, 4 3, 5
2, 1, 2 2, 5 3, 4
3, 1, 3 2, 4 3, 5
2,
1, 1, 4 2, 2 3, 5
2, 1, 4 2, 3 3, 5
3, 1, 2 2, 3 3, 4
3,
1, 1, 4 2, 3 3, 5
2, 1, 4 2, 4 3, 3
3, 1, 5 2, 4 3, 3
```

CROSSED.SPE

```
; FACFORM
; from Keyword file: crossed.key
Facets = 3
Title = Crossed Design
Output = crossed.out
Models = ?,?,?,R5
Data file = crossed.fac
Labels =
1,judge                ; LABELS FOR JUDGES WERE ADDED
1= Judge1
2= Judge2
3= Judge3
*
2,student              ; LABELS FOR STUDENTS WERE ADDED
```

(Appendix continued on next page)

(Appendix continued from previous page)

```
1= Student1
2= Student2
3= Student3
*
3,task                  ; LABELS FOR TASKS WERE ADDED
1=A
2=B
3=C
*
```

CROSSED.FAC

```
1,1,1-3,3,4,5
1,2,1-3,2,5,4
1,3,1-3,3,4,5
2,1,1-3,4,2,5
2,2,1-3,4,3,5
2,3,1-3,2,3,4
3,1,1-3,4,3,5
3,2,1-3,4,4,3
3,3,1-3,5,4,3
```

NESTED DESIGN

NESTED.KEY

```
; File: nested.key
; This program creates a facet specification file and
comma separated data file

$Input    = nested.asc  ; ascii data file
$Output   = nested.fac  ; comma separated data file
$Spoutput = nested.spe  ; facet specifications file

$Facets = 3    ; judge, student, and task (nested design)

; labels for the Facets

$Flabel = 1, "judge"              ;  3 judges
$Flabel = 2, "student"            ;  3 students
$Flabel = 3, "task"               ;  3 tasks

; FACETS specifications

$Spec = "Title  = Nested Design"
$Spec = "Output = nested.out "    ;the FACETS output file
```

(Appendix continued on next page)

(Appendix continued from previous page)

```
$Spec = "Models = ?,?,?,R5"       ; 3 facets using 5 point
rating scale

; read judge number
check1 = $C1
judge=1

$DO = (check1 <>"") ; perform $DO while judges exist
$Label = 1,judge    ; first comma separated value

; skip to next line
$Nextline

; read student number
check2 = $C2
student=1

$DO = (check2 <>""); perform $DO while students exist
$Label = 2,student ; first comma separated value on next
line

; read in three tasks and associated ratings
  $Label=3,$S4W1       ; Task A - col 4
  $Rating = $S7W1      ; rating - col 7
  $Label=3,$S9W1       ; Task B - col 9
  $Rating = $S12W1     ; rating - col 12
  $Label=3,$S14W1      ; Task C - col 14
  $Rating = $S17W1     ; rating - col 17
$Nextline
  check2=$C2
  student=student+1
$Again
  check1 = $C1
  judge=judge + 1
$Again
```

NESTED.ASC

```
1,
1, 1, 3
2, 1, 2
3, 1, 3
2,
1,  2, 2
2,  2, 3
```

(Appendix continued on next page)

(Appendix continued from previous page)

```
3,  2, 3
3,
1,          3, 5
2,          3, 3
3,          3, 3
```

NESTED.SPE

```
; FACFORM
; from Keyword file: nested.key
Facets = 3
Title  = Nested Design
Output = nested.out
Models = ?,?,?,R5
Data file = nested.fac
Labels =
1,judge                        ; LABELS FOR JUDGES WERE ADDED
1=Judge1
2=Judge2
3=Judge3
*
2,student                     ; LABELS FOR STUDENTS WERE ADDED

1=Student1
2=Student2
3=Student3
*
3,task                        ; LABELS FOR TASKS WERE ADDED
1=A
2=B
3=C
*
```

NESTED.FAC

```
1,1-3,1,3,2,3
2,1-3,2,2,3,3
3,1-3,3,5,3,3
```

MIXED DESIGN

MIXED.KEY

```
; File: mixed.key
; This program creates a facet specification file and
```

(Appendix continued on next page)

(Appendix continued from previous page)

```
comma separated data file

$Input    = mixed.asc  ; ascii data file
$Output   = mixed.fac  ; comma separated data file
$Spoutput = mixed.spe  ; facet specifications file

$Facets = 3    ; judge, student, and task (mixed design)

; labels for the Facets

$Flabel = 1, "judge"              ;  3 judges
$Flabel = 2, "student"            ;  3 students
$Flabel = 3, "task"               ;  3 tasks

; FACETS specifications

$Spec = "Title  = Mixed Design"
$Spec = "Output = mixed.out  "    ;the FACETS output file
$Spec = "Models = ?,?,?,R5"       ; 3 facets using 5 point
rating scale

; read judge number
check1 = $C1
judge=1

$DO = (check1 <>"") ; perform $DO while judges exist
$Label = 1,judge    ; first comma separated value

; skip to next line
$Nextline

; read student number
check2 = $C2
student=1

$DO = (check2 <>""); perform $DO while students exist
$Label = 2,student ; first comma separated value on next
line

; read in three tasks and associated ratings
  $Label=3,$S4W1       ; Task A - col 4
  $Rating = $S7W1      ; rating - col 7
  $Label=3,$S9W1       ; Task B - col 9
  $Rating = $S12W1     ; rating - col 12
  $Label=3,$S14W1      ; Task C - col 14
```

(Appendix continued on next page)

(Appendix continued from previous page)

```
  $Rating = $S17W1     ; rating - col 17
$Nextline
  check2=$C2
  student=student+1
$Again
  check1 = $C1
  judge=judge + 1
$Again
```

MIXED.ASC

```
1,
1, 1, 3
2, 1, 2
3, 1, 3
2,
1,       2, 2
2,       2, 3
3,       2, 3
3,
1,          3, 5
2,          3, 3
3,          3, 3
4,
1, 1, 2 2, 4 3, 5
2, 1, 2 2, 3 3, 5
3, 1, 3 2, 4 3, 4
```

MIXED.SPE

```
; FACFORM
; from Keyword file: mixed.key
Facets = 3
Title  = Mixed Design
Output = mixed.out
Models = ?,?,?,R5
Data file = mixed.fac
Labels =
1,judge              ; LABELS FOR JUDGES WERE ADDED
1=Judge1
2=Judge2
3=Judge3
4=Judge4
*
2,student            ; LABELS FOR STUDENTS WERE ADDED
```

(Appendix continued on next page)

(Appendix continued from previous page)

```
1=Student1
2=Student2
3=Student3
*
3,task                ; LABELS FOR TASKS WERE ADDED
1=A
2=B
3=C
*
```

MIXED.FAC

```
1,1-3,1,3,2,3
2,1-3,2,2,3,3
3,1-3,3,5,3,3
4,1,1-3,2,4,5
4,2,1-3,2,3,5
4,3,1-3,3,4,4
```

JOURNAL OF OUTCOME MEASUREMENT®, *3*(4), 339–359
Copyright© 1999, Rehabilitation Foundation, Inc.

Does the Functional Assessment Measure (FAM) Extend the Functional Independence Measure (FIM™) Instrument? A Rasch Analysis of Stroke Inpatients

Richard T. Linn
State University of New York at Buffalo, Buffalo, NY

Richard S. Blair
Sisters of Charity Health Service, Ottawa, Ontario

Carl V. Granger
State University of New York at Buffalo, Buffalo, NY

Dan W. Harper

Patricia A. O'Hara

Edith Maciura
Sisters of Charity Health Service, Ottawa, Ontario

Adding the items of the Functional Assessment Measure (FAM) to the Functional Independence Measure (FIM™ instrument) has been proposed as a method to extend the range of the FIM, particularly when assessing functional status in rehabilitation patients with brain injury, including stroke. It has been proposed that this approach is especially helpful in ameliorating ceiling effects when brain-injured patients have reached the end of their inpatient rehabilitation stay or are being seen in outpatient settings. In the present study, 376 consecutive stroke patients on a Canadian inpatient rehabilitation unit were concurrently administered the FIM and the FAM. Rasch analysis was used to evaluate how well the FAM items extended the difficulty range of the FIM for both the Motor and Cognitive domains. Within the Motor domain, only the FAM item assessing Community Access was found to be more difficult than extant FIM items, and this item showed some tendency to misfit with the other motor items. In the Cognitive domain, the only FAM item with a higher difficulty level than the FIM items was that assessing Employability. Notably, strict adherence to scoring guidelines for these two FAM items requires taking patients out into the community to evaluate their actual performances, a practice unlikely in the typical inpatient stroke rehabilitation unit. Results indicate that use of the entire FAM as an adjunct to the FIM reduces test efficiency while providing only minimal additional protection against ceiling effects.

Requests for reprints should be sent to Richard Linn, Center for Functional Assessment Research, 232 Parker Hall, State University of New York at Buffalo, 3435 Main Street, Buffalo, New York, 14214, e-mail: Rlinn@acsu.buffalo.edu.

The Functional Independence Measure (FIM™) is an 18-item, 7-level scale that reliably and validly assesses change in functional status and burden of care in individuals receiving inpatient rehabilitation (Deutsch, Braun, and Granger, 1997). The FIM is a generic functional assessment tool in the sense that it can be used effectively across a wide spectrum of rehabilitation diagnoses, including but not limited to stroke, spinal cord injury, brain injury, and orthopedic conditions. The FIM has been shown to have predictive validity in terms of predicting minutes of assistance and/or supervision needed by individuals with brain injury (Corrigan, Smith-Knapp, and Granger, 1997; Granger, Divan and Fiedler, 1995).

Concerns have been raised about use of the FIM for individuals receiving rehabilitation for traumatic brain injury because at discharge from the inpatient rehabilitation setting FIM scores tend to be high, suggesting a "ceiling effect" (Hall, Mann, High, Wright, Kreutzer, and Wood, 1996). To address these concerns, Hall and her colleagues developed the Functional Assessment Measure (FAM; Hall, 1997), incorporating all 18 FIM items as well as 12 additional items that assess areas of function not covered by the FIM. The FAM's primary purpose was to improve the sensitivity of the FIM by extending its range of difficulty (FAM Resource Guide, version 6.96; 1996).

The FAM was initially developed for use both with survivors of stroke and traumatic brain injury (Hall, 1992). There are, however, few studies reporting on the use of the FAM with stroke patients. In the single published study known to us to use the FAM to evaluate stroke patients (McPherson, Pentland, Cudmore, and Prescott, 1996), a small mixed sample of neurorehabilitation inpatients containing some individuals with "hemorrhagic brain injury" were assessed with the combined FIM and FAM instruments (FIM+FAM); inter-rater reliability of the FIM+FAM was shown to be good (Kappa values of .50 to .95) for all but one of the FAM items (Adjustment to Limitations). The McPherson, et al., (1996) paper did not address the issue of how well the FAM extends the range of the FIM.

Although the FAM was developed to extend the range of the FIM, its ability to ameliorate ceiling effects remains unknown. Addition of new items to an established instrument, even items with good face validity, may change the internal consistency of the existing test, and thereby affect predictive, content, and/or construct validity. Further, adding items to an existing measure increases the time necessary for data collection,

placing an increased burden upon the patient, the clinician, and ultimately the health care delivery system. Any change in test efficiency, defined here as the ratio between the amount of information provided by new test items and the cost of collecting extra information, produced by addition of test items must be considered before adding new items to an established instrument.

In the current study, all 30 items from the FIM+FAM were administered prospectively to a large group of individuals receiving inpatient rehabilitation services in Canada following stroke. The primary purpose of this study was to compare the scaling characteristics of the FIM and FAM in an inpatient stroke sample. Rasch analysis (Wright and Stone, 1979) was used to determine how well FIM and FAM items measured the same underlying trait, and also to investigate whether FAM items extended the difficulty range of the FIM and thus ameliorated potential ceiling effects.

Methods

Subjects

Participants in this study were selected from the 386 consecutive admissions to the Stroke Rehabilitation Program at Sisters of Charity of Ottawa Health Service (SCOHS) in Ottawa, Ontario between July 1, 1992 to December 31, 1995. Since it was anticipated that the majority of stroke cases would have clear evidence of unilateral dysfunction (i.e., right body dysfunction vs. left body dysfunction), those with bilateral impairment (N = 7) or without overt hemiplegia/hemiparesis (N = 3) were excluded from further analyses. Of the remaining 376 patients, 187 had right-sided hemiplegia/hemiparesis (Right Body Dysfunction group; RBD) and 189 had left-sided hemiplegia/hemiparesis (Left Body Dysfunction group; LBD).

Instruments

The Functional Independence Measure (FIM™) instrument is an 18-item ordinal scale that is used to describe an individual's independence in performing functional activities required to support basic aspects of daily living (Deutsch, Braun and Granger, 1997). Each of the 18 items is evaluated and scored on the same 7-level scale, ranging from a score of "1" (Total Assistance) to "7" (Complete Independence). FIM total scores, derived from summing the 18 individual items, range from 18 to 126; a total score of 126 represents the highest level of independence (Guide for the Uniform Data Set for Medical Rehabilitation, Version 4.0, 1993). Psy-

chometric studies have indicated that the FIM is best represented as two subscales: a 13-item Motor FIM score and a 5-item Cognitive FIM score (Fiedler and Granger, 1996). Previous work has shown that the FIM instrument has good reliability, as well as high face validity, construct validity and predictive validity (Deutsch, Braun and Granger, 1997).

The Functional Assessment Measure (FAM) was developed at the Santa Clara Valley Medical Center by Hall and associates (Hall, 1997) and was designed as an adjunct to the FIM for use with patients with brain injury or stroke. The FAM is meant to address areas of functioning that are not broadly represented in the FIM, such as communication, psychosocial adjustment and cognitive functions (FAM Resource Guide, 6.96, 1996). The FAM consists of 12 items: three items assess the Motor domain, and nine items assess the Cognitive domain. These items are scored on the same seven-point ordinal scale as the FIM. FAM items are rated in a similar time frame as the FIM and, like the FIM, ratings are expected to reflect actual performance rather than the patient's potential capacity to perform an item.

Procedure

Data Collection and Analyses: FIM data were collected using Uniform Data Set for Medical Rehabilitation (UDSMR) standards (Guide for Uniform Data Set for Medical Rehabilitation, version 4.0, 1993); FAM data were collected in accordance with published criteria (FAM Resource Guide, 6.96, 1996). Both sets of standards include collecting the data within 72 hours of admission, and again within 72 hours prior to discharge. Wherever possible, patient's actual functional performances, rather than their potential capacity to perform the function, were assessed. For FIM items, raters scored an item as "1", reflecting Total Dependence, whenever a particular item could not be directly assessed. This practice is in accordance with UDSMR scoring guidelines. However, strict interpretation of scoring guidelines for some FAM items (i.e., Community Access, Employability) requires evaluation of the patient in the community rather than in the hospital, a process not usually feasible for stroke inpatients. For most cases, these items were scored as estimates of the patient's ability, based upon all available information. It is our belief that this scoring approach reflects the method employed in the typical clinical setting. Individuals collecting these data had previous experience administering the FIM and FAM but had not undergone the credentialing process offered by UDSMR. Data were collected by all members of the

rehabilitation team, including physicians, nurses, physiotherapists, occupational therapists, speech and language pathologists, and psychologists. When disagreement regarding scoring individual items arose, the relevant FIM/FAM scoring guidelines were reviewed and team consensus was used to score the item.

When scoring the Locomotion: Walk/Wheelchair FIM item, we applied the scoring standards as set by the UDSMR. In general, scoring for this item ensured that the same mode of locomotion (either walking or using a wheelchair) was used for scoring purposes at both admission and discharge. In cases where the mode of locomotion changed from admission to discharge, both the discharge and admission scores reflected the mode of locomotion being used most frequently at discharge.

Data analyzed in the current study were gathered in Ottawa, Canada. To address potential differences in rehabilitation practices between Canada and the U.S., data collected in Canada were compared to data collected during the same time period by the Uniform Data System for Medical Rehabilitation (UDSMR[SM]) and FIM System[SM]. The FIM System[SM] is used to evaluate outcomes in more than 1,200 rehabilitation facilities in the U.S., as well as in thirteen other countries (UDSMR personal communication, 1998). Demographic and aggregate raw score FIM data from Canada were compared to similar UDSMR data drawn from stroke cases discharged from U.S. facilities in 1994 (Fiedler, Granger and Ottenbacher, 1996). To compare the Rasch-converted measures, a sample of slightly less than 50,000 cases, containing approximately equal numbers of individuals with right and left body dysfunction, was abstracted from the UDSMR dataset, with the prerequisite that this sample include only patients in their first rehabilitation admission for stroke, discharged during calendar year 1994, and with clear evidence of unilateral motor/sensory involvement. Based upon previous experience with Rasch analysis of the UDSMR data, it was anticipated that item difficulty patterns might differ between LBD and RBD patients. Accordingly, these stroke subgroups were analyzed separately.

Data were entered into the SPSS/PC+, version 7.5 Data Entry module (SPSS Base 7.5 Applications Guide, 1997). SPSS was used to analyze demographic trends and to compute measures of central tendency using raw FIM and FAM item scores. In previous, unpublished work, we have found that some of the scaling characteristics of the FIM differ by side of body affected by stroke. To evaluate these differences, demographic characteristics and FIM summary scores for stroke patients with

right body dysfunction (RBD) were compared to stroke patients with left body dysfunction (LBD) using independent sample t-tests.

The BIGSTEPS computer program, version 2.7 (Linacre and Wright, 1997) was used to perform Rasch analysis of FIM and FAM items. Since slightly different scaling characteristics were expected for RBD vs. LBD stroke patients, Rasch analyses were run independently for these two groups and the results are reported separately. The Andrich "Rating Scale" model was applied for all Rasch analyses. This model, which is the default for model selection within the BIGSTEPS program, assumes that the same rating scale structure is shared among the entire set of test items (Wright, 1998). Given that the FIM and the FAM are all scored using the same 1 through 7 scoring levels, and having no reason to suspect differences in scaling among the items, we chose the rating scale model for the present analyses. The BIGSTEPS program converts raw item scores into units known as "logits" using a maximum likelihood estimation process. Person and item measures are estimated simultaneously using the UCON algorithm. This process converts the unequal interval scores present within the FIM+FAM ordinal items into measures with equal intervals (Wright and Masters, 1982). These logit scores were further transformed using the USCALE and UMEAN commands available within BIGSTEPS so that the lowest reportable person measure was 0 and the highest reportable person measure was 100. Converted measure scores were then used to evaluate item difficulty. The BIGSTEPS program offers the potential to "anchor" items from one test form with rasch calibrations from another test form containing overlapping items with the first. We chose not to anchor the items in the current study since we were interested in evaluating the difficulty hierarchy of FIM+FAM items independently for the RBD and LBD groups, as well as for admission and discharge evaluations.

The BIGSTEPS program provides a number of quality control statistics known as "fit" statistics, including 1) Infit Mean Square—reflecting the presence of unexpected outliers or unusual responses in terms of the central range of the measure or person; and 2) Outfit Mean Square—again reflecting unexpected outliers or unusual responses (Lineacre and Wright, 1997). Mean square outfit or infit scores greater than or equal to 1.5 were considered to indicate possible misfit between the specific item and the measurement model reflected by the other items in the scale.

There are no "hard and fast" rules when it comes to setting limits on the size of mean square summary statistics (Wright and Linacre, 1994),

and setting such limits has been a contentious issue in the measurement literature. The higher one sets the upper limit, the greater the likelihood that one will accept items or data with more noise than the proposed model. On the other hand, setting these thresholds too low will cause the possible unnecessary rejection of items or persons. We have established an upper threshold of 1.5 in our laboratory, using rules of thumb proposed by Wright and Linacre (1994), who indicated that acceptable ranges for fit statistics for rating scales are 0.6–1.4 and for clinical observations are 0.5–1.7. Our choice of 1.5 as the upper cut-off represents our interpolation of these rules of thumb, and corresponds closely to recent discussions of evaluations of fit statistics (Linacre, 1999; Smith, 1996). Further, setting thresholds at 1.5 or even higher has support in the recent stroke literature. For example, Segal, Heinemann, Schall and Wright (1997) referenced item fit residual thresholds of > 2.0 and < -2.0 as indicating problems in their evaluations of long-term outcomes following stroke.

Results

Demographic Comparisons

Demographic breakdown of the Canadian stroke sample by age, gender, length of stay, onset to rehabilitation, and side of motor dysfunction is presented in Table 1. Independent groups t-tests comparing the LBD to the RBD patients on continuous demographic data (age, onset to rehabilitation,

Table 1

*Demographic Characteristics of the SCOHS Stroke Sample in Comparison to First Admissions for Stroke from the UDS*MR Database, 1994

Demographic Variable	Left Body Dysfunction	Right Body Dysfunction	UDSMR Stroke Cases–First Admissions 1994
Sample Size	189	187	47,124
Age - Mean years (SD)	67.5 (10.6)	66.6 (11.6)	70.0
Gender (% Male)	49.7	52.9	48.0
FIM Total Score–Admission Mean Score (SD)	84.6 (22.0)	86.1 (24.2)	62.5
FIM Total Score–Discharge Mean Score (SD)	98.9 (23.2)	102.7 (20.7)	86.6
Length of Stay–Mean number of days (SD)	72.1 (25.5)	74.6 (27.4)	25.0
Onset to Rehabilitation Mean number of days (SD)	76.7 (50.8)	70.8 (40.3)	17.0

length of stay) disclosed no significant group differences at the $p < 0.05$ level. Chi-square analysis indicated that the two stroke groups did not differ with respect to the proportion of males to females. T-tests disclosed no differences between the two stroke groups (RBD vs. LBD) with respect to raw (i.e., non-Rasch transformed) FIM scores at admission and raw FIM scores at discharge. These findings suggest that on average the two groups were of equal severity both at the beginning and at the end of inpatient rehabilitation.

Non-statistical comparison of the Canadian stroke sample to the 1994 UDSMR data (See Table 1) suggested that on average Canadian patients were less disabled than U.S. patients upon admission to rehabilitation. For example, the Canadian patients remained in the acute setting before being transferred to rehabilitation more than four times as long as their U.S. counterparts (See Table 1). Canadian stroke patients had rehabilitation lengths of stay that were almost three times higher, and had higher admission and discharge FIM scores, than patients in the U.S. sample. The higher admission FIM scores in the Canadian sample probably reflect the increased length of time Canadian patients spend in acute care: the longer these patients remain in acute care, the more functional recovery they are likely to show prior to entering the rehabilitation setting. Higher FIM scores at discharge in the Canadian sample are pertinent to the present analyses since one stated purpose of developing the FAM was to help overcome potential ceiling effects in the FIM. Such ceiling effects should be more apparent in a stroke sample with higher general functional status than is typical in the U.S. during inpatient rehabilitation.

Analysis of Floor and Ceiling Effects

To evaluate the presence of floor and ceiling effects in the Canadian sample, the frequency distributions of raw FIM and FIM+FAM summary scores for three indices (Total Score, Motor Score and Cognitive Score) were evaluated. Table 2 displays the results of these analyses. To evaluate ceiling effects, we assessed the extent to which summary scores at admission or discharge were scored at an average item score of 6.0 or higher (i.e., Modified Independence level). A threshold of 6.0 was used as a marker of potential ceiling effects because previous work with the FIM has shown that the burden of care placed upon a care provider does not differ between patients with average item scores of 6.0 and patients with the highest possible average item score (i.e., 7.0). The frequency analysis shown in Table 2 demonstrates that potential ceiling effects were

Table 2

Summary of Ceiling and Floor Effects for FIM and the FIM+FAM raw summary scores

Ceiling Effects	**Percent of patients with average item scores of 6.0 or greater at admission**	**Percent of patients with average item scores of 6.0 or greater at discharge**
FIM Motor Score	23.9	54.0
FIM Cognitive Score	29.8	41.5
FIM Total Score	18.4	52.9
FIM+FAM Motor Score	22.6	53.7
FIM+FAM Cognitive Score	21.3	39.6
FIM+FAM Total Score	13.8	41.8
Floor Effects	**Percent of patients with average item scores of less than 2.0 at admission**	**Percent of patients with average item scores of less than 2.0 at discharge**
FIM Motor Score	5.1	2.1
FIM Cognitive Score	1.6	1.1
FIM Total Score	2.1	.5
FIM+FAM Motor Score	4.0	1.3
FIM+FAM Cognitive Score	1.6	1.1
FIM+FAM Total Score	1.6	.5

present in the FIM in 18–30% of patients at admission and in 40–54% of patients at discharge. When FIM+FAM scores were considered using the same criteria, between 14–23% of patients demonstrated potential ceiling effects at admission, and 40–54% of patients demonstrated potential ceiling effects at discharge.

Table 2 demonstrates that floor effects were not a major problem within this data set for either the FIM alone or for the FIM+FAM combined scale. Using the criterion of average item scores of less than 2.0, fewer than 5% of patients demonstrated potential floor effects at either admission or discharge.

Rasch Analysis of FIM Motor Items

Fit statistics from Rasch analysis of FIM Motor items for the RBD stroke patients are presented in Table 3, separated by admission and discharge. Table 3 shows that the Eating, Bowel, Bladder, and, to a lessor extent, Locomotion: Walk/Wheelchair items were misfitting in the Ottawa data, both at admission and discharge. Table 3 provides a similar evaluation of Rasch fit statistics for LBD stroke cases. For this subgroup of stroke patients, the Eating, Bowel, Bladder and Locomotion: Walk/Wheelchair items were again misfitting.

Table 3

Rasch Mean Square Infit and Outfit Statistics for FIM Motor Items for SCOHS Stroke Cases, Categorized by Side of Body Dysfunction

Item	Right Body Dysfunction				Left Body Dysfunction			
	Infit Admission	Outfit Admission	Infit Discharge	Outfit Discharge	Infit Admission	Outfit Admission	Infit Discharge	Outfit Discharge
Eating	2.71*	5.34*	3.40*	9.90*	2.45*	9.90*	2.63*	9.90*
Grooming	1.21	0.90	1.41	0.86	1.20	0.99	1.12	0.61
Bowel	1.81*	3.10*	1.86*	2.42*	1.29	2.08*	1.52*	2.46*
Bladder	2.42*	2.63*	2.26*	1.36	2.09*	1.58*	2.72*	2.24*
Dress–Upper	0.93	1.00	1.42	0.84	1.05	1.12	0.98	0.85
Transfer: Bed	0.48	0.58	0.57	0.49	0.37	0.44	0.37	0.40
Transfer: Toilet	1.07	0.87	0.42	0.39	0.32	0.39	0.35	0.38
Toileting	0.51	0.45	1.01	0.66	0.46	0.46	0.67	0.54
Bathing	0.40	0.60	0.79	0.69	0.43	0.58	0.55	0.61
Dressing–Lower	0.48	0.51	0.97	0.78	0.59	0.54	0.78	0.62
Locomotion: Walk/Wheelchair	1.78*	2.03*	1.59*	1.50*	2.37*	2.45*	1.88*	2.17*
Transfer: Tub	1.07	0.87	0.80	0.73	0.73	0.57	0.65	0.60
Locomotion: Stairs	1.00	0.91	0.79	0.85	0.95	0.70	0.91	0.70

*Indicates fit statistic scores ≥ 1.5 suggesting possible misfit

Rasch Analysis of FIM Cognition Items

Rasch fit statistics for the FIM Cognition items are shown in Table 4. For RBD cases the Social Interaction and Expression items were misfitting, whereas for LBD cases none of the items were misfitting.

Table 4

Rasch Mean Square Infit and Outfit Statistics for FIM Cognition Items for SCOHS Stroke Cases, Categorized by Side of Body Dysfunction

Item	Right Body Dysfunction				Left Body Dysfunction			
	Infit Admission	Outfit Admission	Infit Discharge	Outfit Discharge	Infit Admission	Outfit Admission	Infit Discharge	Outfit Discharge
Problem Solving	0.66	0.68	0.65	0.65	0.76	0.73	0.78	0.75
Memory	0.90	1.08	0.76	0.89	1.35	1.30	1.39	1.28
Social Interaction	1.30	1.47	1.29	1.72*	1.05	1.15	0.82	1.00
Comprehension	1.12	1.17	1.23	1.40	0.80	0.68	0.77	0.67
Expression	2.08*	1.75*	1.94*	1.72*	1.16	0.96	1.27	1.04

*Indicates fit statistic scores ≥ 1.5 suggesting possible misfit

Rasch Analysis of FIM+FAM Motor Items

Rasch fit statistics for the combined FIM+FAM Motor items from the Canadian sample are shown in Tables 5 and 6. For LBD cases, the Swallowing and Eating items were relatively severely misfitting, the Bowel and Bladder items were moderately misfitting, and the Locomotion: Walk/Wheelchair item was marginally misfitting. For individuals with RBD, Eating, Swallowing, Bowel, Bladder and Community Access items were

misfitting. The FIM Locomotion: Walk/Wheelchair item was misfitting at admission, but was not misfitting at discharge.

Table 5

Mean Square Infit and Outfit Statistics for the SCOHS Data, FIM & FAM Motor Items Combined, For Left Body Dysfunction

	SCOHS Data			
	Admission		Discharge	
Item	Infit	Outfit	Infit	Outfit
Locomotion: Stairs	0.95	0.70	0.80	0.60
Community Access*	1.12	1.27	1.23	1.33
Transfer–Tub	0.72	0.57	0.57	0.52
Transfer–Car*	0.41	0.38	0.57	0.56
Locomotion: Walk/Wheelchair	2.32**	2.45**	1.86**	2.07**
Dressing–Lower	0.62	0.57	0.90	0.70
Bathing	0.43	0.51	0.58	0.58
Toileting	0.49	0.48	0.77	0.61
Transfer–Toilet	0.31	0.35	0.37	0.38
Transfer–Bed	0.35	0.39	0.38	0.37
Dressing–Upper	1.06	1.09	1.05	0.86
Bladder	2.01**	1.56**	2.70**	2.21**
Bowel	1.23	1.77**	1.56**	2.59**
Grooming	1.20	0.95	1.11	0.60
Swallowing*	1.93**	7.96**	1.88**	7.58**
Eating	2.32**	9.90**	2.49**	9.90**

*Indicates FAM item

**Indicates fit statistic scores ≥ 1.5 suggesting possible misfit

Rasch item difficulty hierarchies for the FIM+FAM Motor items are depicted in Table 7. For RBD patients evaluated at discharge, the Community Access item was more difficult than the remaining items. Otherwise, the Community Access item was easier than the FIM Stairs item. The Swallowing item, on the other hand, was generally easier than the majority of FIM or FAM Motor items, but notably was not easier than the FIM Eating item. The Car Transfer item was moderately difficult, but less difficult than the FIM Transfer: Tub motor items.

Rasch Analysis of FIM + FAM Cognition Items

Table 8 shows Rasch fit statistics for individuals with LBD. The Speech Intelligibility item was most misfitting, and the Writing, Expression and Orientation items were also misfitting. Similar fit statistics for RBD patients are provided in Table 9. The Orientation item was marginally misfitting for RBD cases at discharge.

Table 6

Mean Square Infit and Outfit Statistics for the SCOHS Data, FIM & FAM Motor Items Combined For Right Body Dysfunction

	SCOHS Data			
	Admission		Discharge	
Item	Infit	Outfit	Infit	Outfit
Locomotion: Stairs	0.90	0.79	0.77	0.75
Community Access*	2.09**	2.08**	1.61**	1.81**
Transfer–Tub	0.91	0.74	0.68	0.62
Transfer–Car*	0.65	0.61	0.65	0.59
Locomotion: Walk/Wheelchair	1.58**	1.61**	1.43	1.29
Dressing–Lower	0.51	0.54	1.01	0.81
Bathing	0.40	0.45	0.79	0.65
Toileting	0.51	0.46	1.00	0.65
Transfer–Toilet	0.42	0.41	0.41	0.38
Transfer–Bed	0.45	0.53	0.54	0.46
Dressing–Upper	0.85	0.88	1.34	0.77
Bladder	2.21**	2.26**	2.13**	1.40
Bowel	1.57**	2.28**	1.67**	1.93**
Grooming	1.10	0.81	1.30	0.81
Swallowing*	1.93**	3.48**	2.60**	4.31**
Eating	2.51**	3.70**	3.27**	6.64**

*Indicates FAM item

**Indicates fit statistic scores ≥ 1.5 suggesting possible misfit

Rasch item difficulty hierarchies for the FIM+FAM Cognition items for LBD cases are shown in Table 10. The Employability item was found to extend the FIM in the difficult direction, but none of the FAM items were found to extend the FIM at the easy end of the difficulty continuum. Table 10 also displays Rasch measure scores for the RBD cases. The Employability item was again found to extend the FIM in the difficult direction, and the Orientation item was found to extend the FIM in the easy direction.

Discussion

In a previous report outlining the results of Rasch analysis of the FAM, Hall, et al., (1993) stated "The FAM items were more widely spread in difficulty than the FIM items (average scores for patients were highest for swallowing and lowest for community mobility, stairs, and car transfers. These FAM items extended the range of measured difficulty beyond that of FIM items" (p. 67). In the present study, Rasch analysis was also

Table 7

Rasch Converted Measure Scores for 13 FIM and 3 FAM Motor Items

	SCOHS Data			
Item	Left Body Dysfunction Admission	Left Body Dysfunction Discharge	Right Body Dysfunction Admission	Right Body Dysfunction Discharge
Locomotion: Stairs	67.0	68.0	64.0	67.0
Community Access*	60.0	65.0	63.0	71.0
Transfer–Tub	64.0	65.0	63.0	63.0
Transfer–Car*	61.0	63.0	60.0	62.0
Locomotion: Walk/Wheelchair	55.0	49.0	55.0	51.0
Dressing–Lower	57.0	56.0	55.0	53.0
Bathing	54.0	55.0	53.0	54.0
Toileting	53.0	54.0	53.0	51.0
Transfer–Toilet	56.0	55.0	54.0	54.0
Transfer–Bed	55.0	53.0	53.0	53.0
Dressing–Upper	51.0	51.0	49.0	44.0
Bladder	42.0	45.0	41.0	42.0
Bowel	37.0	40.0	40.0	46.0
Grooming	38.0	38.0	41.0	40.0
Swallowing*	28.0	24.0	35.0	30.0
Eating	22.0	18.0	23.0	19.0

Note: Higher measure scores indicate items of greater difficulty
*Indicates FAM item

Table 8

Mean Square Infit and Outfit Statistics for the SCOHS Data, FIM & FAM Cognitive Items Combined For Left Body Dysfunction

	SCOHS Data			
	Admission		Discharge	
Item	Infit	Outfit	Infit	Outfit
Employability*	0.79	0.81	0.76	0.73
Problem Solving	0.87	0.92	0.86	0.86
Safety Judgement*	0.55	0.64	0.47	0.47
Adjustment to Limitations*	0.61	0.67	0.77	0.76
Memory	0.99	1.02	0.93	0.93
Emotional Status*	0.98	1.10	1.14	1.21
Attention*	1.08	1.04	0.99	0.91
Orientation*	1.32	1.08	1.65**	1.05
Social Interaction	1.06	1.03	0.96	0.97
Writing*	1.65**	1.45	1.87**	1.81**
Reading*	1.05	0.94	1.32	1.34
Speech Intelligibility*	1.93**	2.27**	1.22	2.24**
Expression	1.80**	1.67**	1.89**	1.74**
Comprehension	1.05	0.69	1.25	0.79

*Indicates FAM item
** Indicates fit statistic score $\geq$ 1.5 suggesting possible misfit

Table 9

Mean Square Infit and Outfit Statistics for the SCOHS Data, FIM & FAM Cognitive Items Combined For Right Body Dysfunction

	SCOHS Data			
	Admission		Discharge	
Item	Infit	Outfit	Infit	Outfit
Employability*	0.81	0.89	0.91	0.88
Problem Solving	0.75	0.76	0.86	0.81
Safety Judgement*	0.63	0.79	0.49	0.64
Adjustment to Limitations*	0.69	0.77	0.65	0.68
Memory	1.37	1.25	1.46	1.28
Emotional Status*	0.81	0.95	0.75	0.93
Attention*	1.34	1.27	1.22	1.01
Orientation*	1.43	1.07	2.04**	1.26
Social Interaction	0.76	0.79	0.82	0.92
Writing*	0.95	0.90	1.06	1.02
Reading*	0.69	0.64	0.79	0.73
Speech Intelligibility*	1.53	1.34	1.71	1.47
Expression	1.33	1.20	1.41	1.16
Comprehension	1.00	0.87	0.95	0.77

*Indicates FAM item
**Indicates fit statistic ≥ 1.5 suggesting possible misfit

Table 10

Rasch Converted Measure Scores for 5 FIM and 9 FAM Items

	Left Body Dysfunction		Right Body Dysfunction	
Item	SCOH Admission	SCOH Discharge	SCOH Admission	SCOH Discharge
Employability*	76.0	75.0	67.0	67.0
Problem Solving	61.0	65.0	56.0	57.0
Safety Judgement*	62.0	61.0	54.0	52.0
Adjustment to Limitations*	58.0	58.0	51.0	52.0
Memory	57.0	59.0	50.0	50.0
Emotional Status*	55.0	55.0	48.0	48.0
Attention*	54.0	55.0	44.0	42.0
Orientation*	49.0	48.0	43.0	38.0
Social Interaction	47.0	45.0	43.0	42.0
Writing*	40.0	42.0	51.0	54.0
Reading*	38.0	39.0	48.0	49.0
Speech Intelligibility*	38.0	36.0	48.0	48.0
Expression	33.0	31.0	51.0	52.0
Comprehension	32.0	31.0	47.0	47.0

Note: Higher logit scores indicate items of greater difficulty
*Indicates FAM item

used to evaluate the degree to which adding items from the FAM increased the FIM's range. Increases in a test's range can be gained through extensions of a scale at either end of the difficulty continuum, so we evaluated how well the FAM items extended the FIM for both the most difficult and the easiest items. However, since the FIM has most often been cited as potentially showing ceiling effects, especially at discharge from rehabilitation (Hall, et al., 1996), we concentrate in the present discussion on how many of the FAM items were more difficult than FIM items.

In addition to evaluating the difficulty range of FIM+FAM items, we also assessed within the Rasch analyses the degree to which FAM items "fit" with FIM items in measuring the same underlying construct. This practice has previously been shown to be an effective method to evaluate the quality of functional assessment instruments (Tesio, Perucca, Battaglia, and Franchignoni, 1997).

Before addressing the scaling characteristics of the FIM+FAM, it is instructive to note that in this Canadian stroke sample the FIM alone demonstrated evidence of ceiling but not floor effects. Ceiling effects were seen as potentially affecting as many as 30% of these inpatients at admission and more than 50% of the inpatients at discharge. The current findings may therefore be seen as supporting the presence of ceiling effects for the FIM in this Canadian inpatient sample. It must be stressed, however, that the Canadian sample had a strikingly different mean functional profile at rehabilitation admission when compared to averages taken from a large U.S. sample. Admission FIM scores were, on average, 22 points higher in the Canadian sample, and discharge FIM scores were 12 points higher. Further analysis would be needed to determine if ceiling effects are present when the FIM is applied to stroke patients in the typical U.S. setting.

Rasch analysis of the FIM+FAM in this stroke sample showed that the only FAM Motor item demonstrating a higher difficulty level than extant FIM Motor items was that measuring Community Access. Of note, the Community Access item surpassed the FIM Stairs item in difficulty level only for stroke patients with right body dysfunction at discharge. The Car Transfer item from the FAM was consistently easier than the FIM Transfer: Tub item. Hall, et al., (1993) previously reported that the FAM Swallowing item was consistently the easiest item in their sample. However, in the current investigation, we found that the FIM Eating item was easier than the Swallowing item both at admission and at discharge.

Rasch analysis of the nine FAM Cognition items indicated that only

one item, Employability, was more difficult than extant FIM Cognition items. At the easy end of the difficulty hierarchy, the FIM Comprehension item was consistently easier than the FAM items. The Employability item also demonstrated good fit statistics, suggesting that this FAM item was measuring a similar underlying trait as that assessed by the other FIM Cognition items. Most of the remaining FAM Cognition items demonstrated acceptable fit statistics, with the exception of the item assessing Speech Intelligibility.

Rasch converted measure scores for each of the FIM+FAM motor items, separated by side of body affected and admission vs. discharge status, are displayed in Figure 1. In this figure, it can be seen that in general, there were very few differences in the item hierarchy rankings when comparing stroke patients with LBD versus those with RBD. The Stairs item tended to be the most difficult item, except for RBD patients at discharge where the Community Access item was most difficult. At the easy end of the difficulty hierarchy, the Bowel item was easier than

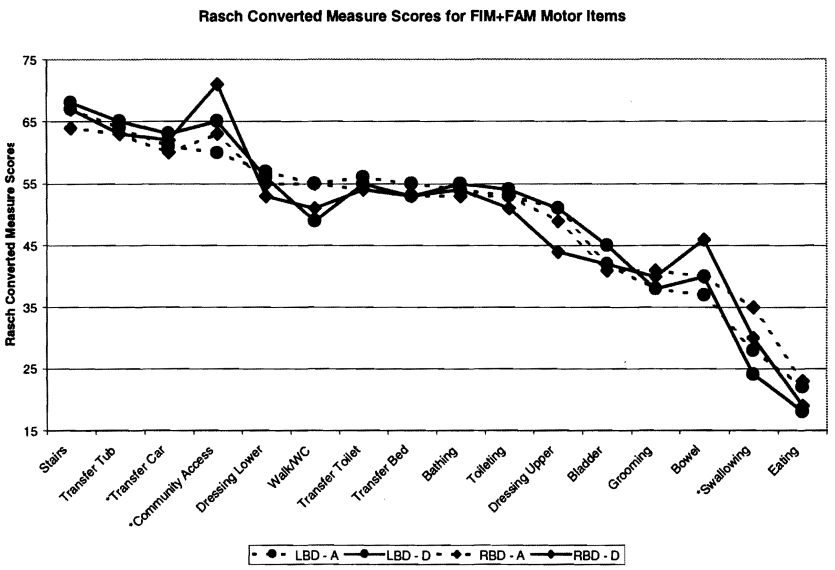


Figure 1. Rasch converted measure scores for the FIM+FAM Motor items, separated by side of body involvement and time of assessment (LBD - A: Left Body Dysfunction at Admission; LBD - D: Left Body Dysfunction at Discharge; RBD - A: Right Body Dysfunction at Admission; RBD-D: Right Body Dysfunction at Discharge). Higher scores represent items of greater difficulty. The three FAM items are identified with an *.

the Bladder item for LBD at both admission and discharge but was more difficult than the Bladder item for RBD at discharge. The Grooming item was easier than the Bowel item for the LBD group but was slightly more difficult than the Bowel item for the RBD group at admission. There was less separation between the Swallowing and Eating items for the LBD group than there was for the RBD group. Comparisons of admission and discharge rankings disclosed negligible differences.

Rasch converted measure scores for each of the FIM+FAM cognitive items, separated by side of body affected and admission vs. discharge status, are displayed in Figure 2. In this figure, differences between the LBD and RBD item hierarchies are clearly evident for items containing measurement of language-related activities (Writing, Reading, Speech Intelligibility, Expression, and Comprehension). Whereas for the LBD group these items tended to be easier than some other cognitive items, for the RBD group these items were more difficult or equally difficult as other cognitive items such as Social Interaction, Orientation, or Attention. Likely, this distinction represents the differential presence of aphasia in the RBD (presumably left-hemisphere involved) group, and possibly the differential presence of attentional and/or orientation problems in the LBD (presumably right-hemisphere involved) group. Within both the LBD or RBD groups, the Employability item emerged consistently as the most difficult item. Shifts in the overall difficulty hierarchy from admission to discharge again appeared negligible.

The findings summarized above provide little justification for use of the FAM *as a whole* to extend the range of the FIM and to decrease potential ceiling effects. Only two of the 12 FAM items demonstrated greater item difficulty than existing FIM items. Therefore, only these two items appear to have utility in ameliorating potential ceiling effects. Using the entire set of items offered within the FAM would needlessly increase test length without appreciably increasing test difficulty, thereby producing a net loss in test efficiency.

One potential outcome from the present investigation is for clinicians to add to the FIM the two FAM items that did extend the difficulty range of the FIM, namely the Community Access and Employability items. However, some concerns must be raised with this approach. First, the Community Access item was most difficult in only one out of four situations evaluated (i.e., when assessing RBD patients at discharge), suggesting that there may be a narrow application window for this item. Second, the Community Access item did not always demonstrate acceptable fit statistics, raising the possibil-

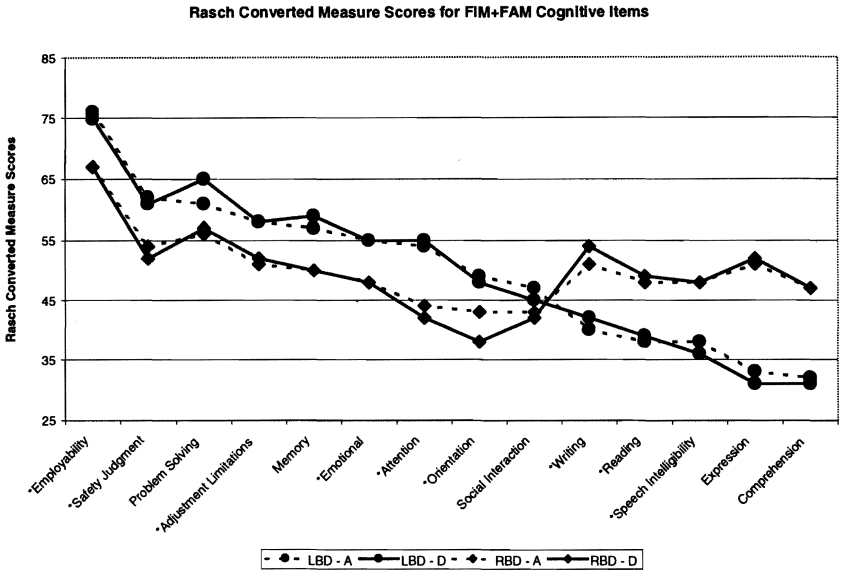


Figure 2. Rasch converted measure scores for the FIM+FAM Cognitive items, separated by side of body involvement and time of assessment (LBD - A: Left Body Dysfunction at Admission; LBD - D: Left Body Dysfunction at Discharge; RBD - A: Right Body Dysfunction at Admission; RBD-D: Right Body Dysfunction at Discharge). Higher scores represent items of greater difficulty. The nine FAM items are identified with an *.

ity that this item addresses a construct that is distinct from that represented by the existing FIM Motor items. An alternate interpretation is that the Community Access does measure the same construct as the FIM Motor items, but was not being scored in a consistent fashion in this sample. Scoring issues are always a concern when behaviors must be estimated rather than observed, a potential problem with both the Community Access and Employability items. Third, although the 18-item FIM has been widely accepted and standardized, no national benchmarks are available for the FIM plus these two FAM items, leaving clinicians with little means to interpret their findings if they did supplement the FIM. Lastly, although these two FAM items do extend the difficulty range of the FIM, the increase in range does not appear to be substantial enough to warrant changing the existing FIM instrument and associated national benchmarks to accommodate the items.

Recent work by Tesio and Cantagallo (1998) is consistent with this latter argument. In their study of outpatients with brain injury, Tesio and Cantagallo (1998) reported that, on average, the FAM items were too easy,

and in general both the FAM and the FIM demonstrated ceiling effects in this population. In their study, the Employability item was also the most difficult Cognitive item, but they noted that it was only slightly more difficult than the most difficult Cognitive item from the FIM.

In the current sample the Eating, Bowel, Bladder and Locomotion: Walk/Wheelchair FIM motor items were misfitting. Previous work in the U.S. using Rasch analysis has also shown that the FIM Bowel and Bladder items tend to misfit (Heinemann, Linacre, Wright, Hamilton, and Granger, 1994). One interpretation for this consistent finding is that the FIM Bowel and Bladder items involve assessment of both the level of assistance needed for sphincter management as well as the presence or absence of continence, leading some investigators to suggest that the items be changed (Heinemann, et al., 1994). The misfit attributable to the Walk/Wheelchair item may stem from the fact that some individuals are scored on this item when using wheelchairs while others are scored without the use of assistive devices. This item may need to be separated into two items. The reason why the Eating item was misfitting in this sample is not clear.

The FAM was designed to help adjust for ceiling effects in the difficulty level of items on the FIM when assessing the functional status of brain injured and stroke patients receiving medical rehabilitation, and also to provide for evaluation of domains of interest not fully represented on the FIM. The current investigation suggests that only two of the FAM items have the potential to extend the difficulty level of the FIM and thus ameliorate ceiling effects, and one of these items was misfitting in the present sample. In general, our findings suggest that the increase in test length and expense associated with adding all of the FAM items to the FIM would not be offset by a substantial increase in the difficulty level of the FIM, thereby producing a net loss in overall test efficiency. The current research does support use of two FAM items in conjunction with the FIM to address potential ceiling effects. However, clinicians may find it burdensome to interpret FIM + 2 FAM scores in the absence of national benchmarks for this combination.

Acknowledgment

This work was supported in part by National Institute of Child Health and Human Development Postdoctoral Fellowship/National Research Service Award #T32HD07423 to Dr. Linn.

References

Corrigan, J. D., Smith-Knapp, K., and Granger, C. V. (1997). Validity of the Functional Independence Measure for Persons with Traumatic Brain Injury. *Archives of Physical Medicine and Rehabilitation, 78*, 828-834.

Deutsch, A., Braun, S., and Granger, C. V. (1997). The Functional Independence Measure (FIM[SM] Instrument). *Journal of Rehabilitation Outcomes Measurement, 1*(2), 67-71.

Fiedler, R. C., and Granger, C.V. (1996). The Functional Independence Measure: A measurement of disability and medical rehabilitation. In N. Chino and J.L. Melvin (Eds.) *Functional Evaluation of Stroke Patients*. Tokyo: Springer Verlag, 75-92.

Fiedler, R. C., Granger, C. V., and Ottenbacher, K. J. (1996). The Uniform Data System for Medical Rehabilitation: Report of First Admissions for 1994. *American Journal of Physical Medicine and Rehabilitation, 75*(2), 125-129.

Functional Assessment Measure Resource Guide, Version 6.96. (1996). Santa Clara, CA: Santa Clara Valley Medical Center.

Granger, C. V., Divan, N., and Fiedler, R. C. (1995). Functional Assessment Scales: A study of persons after Traumatic Brain Injury. *American Journal of Physical Medicine and Rehabilitation, 74*(2), 107-113.

Guide for the Uniform Data Set for Medical Rehabilitation, Version 4.0. (1993). Buffalo, NY: State University of New York at Buffalo.

Hall, K. (1997). The Functional Assessment Measure (FAM). *Journal of Rehabilitation Outcomes Measurement, 1*(3), 63-65.

Hall, K. M. (1992). Overview of functional assessment scales in brain injury rehabilitation. *NeuroRehabilitation, 2*(4), 98-113.

Hall, K. M., Hamilton, B. B., Gordon, W. A., and Zasler, N. D. (1993). Characteristics and comparisons of functional assessment indices: Disability Rating Scale, Functional Independence Measure, and Functional Assessment Measure. *Journal of Head Trauma and Rehabilitation, 8*(2), 60-74.

Hall, K. M., Mann, N., High, W., Wright, J., Kreutzer, J., and Wood, D. (1996). Functional measures after traumatic brain injury: Ceiling effects of FIM, FIM+FAM, DRS and CIQ. *Journal of Head Trauma and Rehabilitation, 11*, 27-39.

Heinemann, A. W., Linacre, J. M., Wright, B. D., Hamilton, B. B., and Granger, C. V. (1994). Measurement characteristics of the Functional Independence Measure. *Topics in Stroke Rehabilitation, 1*(3), 1-15.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.

Linacre, J. M., and Wright, B. D. (1997). *A user's guide to BIGSTEPS: Rasch-Model Computer Program.* Chicago: MESA.

McPherson, K. M., Pentland, B., Cudmore, S. F., and Prescott, R. J. (1996). An inter-rater reliability study of the Functional Assessment Measure (FIM + FAM). *Disability and Rehabilitation, 18*(7), 41-347.

Personal Communication (September, 1998). Uniform Data System for Medical Rehabilitation.

SPSS® Base 7.5 Applications Guide (1997) Chicago: SPSS, Inc.

Segal, M. E., Heinemann, A. W., Schall, R. R., and Wright, B. D. (1997). Rasch analysis of a brief physical ability scale for long-term outcomes of stroke. In R. Smith (Ed.) *Physical Medicine and Rehabilitation: State of the Art Reviews, 11*(2), 385-396.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions, 10*(3), 516-517.

Tesio, L., and Cantagallo, A. (1998). The Functional Assessment Measure (FAM) in Closed Traumatic Brain Injury outpatients: A Rasch-based psychometric study. *Journal of Outcome Measurement, 2*(2), 79-96.

Tesio, L., Perucca, L., Battaglia, M. A., and Franchignoni, F. (1997). Quality assessment of the FIM (Functional Independence Measure) ratings through Rasch analysis. *Europa Medicophysica, 33*, 69-78.

Wright, B. D. (1998). Model selection: Rating scale or partial credit? *Rasch Measurement Transactions, 12*(3), 641-642.

Wright, B. D., and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright B. D., and Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

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

JOURNAL OF OUTCOME MEASUREMENT®, 3(4), 360–381
Copyright© 1999, Rehabilitation Foundation, Inc.

Measuring Change across Multiple Occasions Using the Rasch Rating Scale Model

Edward W. Wolfe
Michigan State University

Chris W. T. Chiu
University of Pittsburgh

When latent trait models are used to measure change across time, it is difficult to disentangle changes in one facet of the measurement context from changes in other facets. Hence, it is difficult to diagnose change. Wright (1999b) proposed an algorithm for disentangling change, and previously the authors applied this algorithm to measuring change across two occasions (Wolfe and Chiu, 1999). In this article, we extend Wright's algorithm to disentangle changes in measures across three occasions. We describe a standard Rasch rating scale analysis of a multi-occasion evaluation that produces confusing results when subjected to a series of "separate" calibrations. Then, we apply Wright's correction to the same data to show that the algorithm reveals changes that are more similar to ones that would be expected. Our demonstration shows that Wright's procedure can reduce misfit to the Rasch Rating Scale Model as well as changing the interpretation of change within the measurement context.

The data upon which this article is based were provided by the Detroit Skillman Parenting Program, Department of Psychiatry at Michigan State University. A previous version of this manuscript was presented in November 1998 at the Florida Educational Research Association, Orlando, FL.
Requests for reprints should be sent to Edward W. Wolfe, 459 Erickson Hall, Michigan State University, East Lansing, MI 48824, e-mail: wolfee@msu.edu.

Evaluating changes in outcome measures over time presents several potential problems. A number of confounds may distort true changes, making it unclear whether observed changes (or lack of changes) are due to the intervention under study or some other uncontrolled design effect. These uncontrolled effects could introduce apparent changes in participants that are untrue. For example, apparent change in participants may be due to unique events that occur during the intervention (e.g., mortality, maturation, history). Alternatively, these uncontrolled effects could obscure changes in participants, implying that there is no change when, indeed, changes did take place. These obscuring effects may be due to confounds that influence evaluative study control groups (e.g., treatment diffusion), the observational procedures (e.g., the Hawthorne effect), or statistical procedures (e.g., unreliability of measurement instruments, regression toward the mean) (Cook and Campbell, 1979).

Those who design studies to evaluate the effectiveness of an intervention typically implement procedures and features in the designs of their studies so that each of these potential confounds to detecting changes in participants is minimized to the greatest extent possible. However, even the most carefully designed study may be subject to potential confounds that are inherent in the very nature of change itself. More specifically, when individuals are measured across multiple occasions using typical psychological or behavioral instruments (i.e., instruments that measure a latent trait by employing several observable indicators of that latent trait), it becomes difficult to determine which facets of the measurement system are changing and which ones are not (Wright, 1996b). For example, the locations of individual people on the underlying continuum that represents the latent trait may shift. On the other hand, if change is measured through self-reports or observational ratings, the raters may change the ways that they define the rating scale categories from one time to another. Additionally, if the intervention is successful, we might expect the indicators that we employ to measure this change (e.g., test items) to change their relative positions on the underlying scale. That is, we might expect individuals not only to change their positions on the underlying continuum as the result of the intervention, but we might also expect individuals to change the ways that they conceptualize or execute the construct being measured (Smith, 1997).

Unfortunately, the models that we use to measure these changes in individuals, rating scales, and items do not disentangle changes in one

facet of the evaluative context from changes in other facets. As a result, we are left with ambiguous diagnosis of the nature of change in our evaluative study. Previously, Wright (1996b) proposed a Rasch-based algorithm for disentangling these changes, and Wolfe and Chiu (1999) demonstrated how this algorithm is applied to evaluative settings involving measurement across two occasions. Many evaluative studies involve measures that are taken across three or more occasions. The purpose of this article is to extend Wright's algorithm by describing how that method can be applied to disentangle changes across three time points. To this end, we describe the results of a standard analysis of a multi-occasion evaluation that produces confusing results that would cause one to draw contradictory conclusions about the intervention in question. We then describe and apply Wright's correction algorithm to the same data to reveal how the expected and predictable changes in the evaluative study were obscured by the spill over of change between the multiple facets of our measurement situation.

Background

To evaluate changes in persons over time, the items and rating scales that are used to measure changes in persons must be stable across multiple administrations of the measurement instrument. Only when items and scales demonstrate invariance across occasions can differences between multiple measures of individual persons be validly interpreted (Smith, 1997; Wilson, 1992; Wright, 1996b). When items and scales do not demonstrate invariance across occasions, the changes in these facets of the measurement context may obscure the true nature of changes in person measures. Additional problems in the interpretation of changes in persons may be introduced when items are added, deleted, or edited from the original measurement instrument; observations are not available for all persons on all items across all time points; or rating scale categories are changed from one time point to another (Roderick and Stone, 1996, April). Because of these issues, scaling methods are often used to place measures from different administrations of a measurement instrument onto a common underlying scale.

Rasch Rating Scale Model

The Rasch Rating Scale Model (RSM) describes the probability that a specific person (n) will be rated on a particular rating scale item (i) with a specific rating scale category (x) (Andrich, 1978). The equation for this

probability (Equation 1) contains three parameters: the person's ability (β_n), the item's difficulty (δ_i), and the category threshold (i.e., the threshold between two adjacent scale levels) (τ_x). In this model, it is assumed that the distance between each category threshold is constant across all items. Calibration of data with this model results in a separate parameter estimate and a standard error for that estimate for each person, item, and category threshold in the measurement context.

$$P(X_{ni} = x) = \frac{\exp \sum_{j=0}^{x} [\beta_n - (\delta_i + \tau_j)]}{\sum_{x=0}^{m} \exp \sum_{j=0}^{x} [\beta_n - (\delta_i + \tau_j)]}, x = 0, 1, \ldots, m \quad (1)$$

where $P(X_{ni} = x)$ is the probability that a person n is assigned to rating scale category x on item i, which has $m + 1$ rating scale categories, and

$$\sum_{j=0}^{0} [\beta_n - (\delta_i + \tau_j)] = 0 .$$

Expected values for the RSM can be calculated for any combination of person ability and item difficulty (Equation 2). Departures in the data from these expected values indicate potential misfit and are captured by a fit statistic associated with each parameter estimate. The mean square **outfit** statistic is based on the mean of the squared standardized residuals of the observed ratings from their expected values (Wright and Masters, 1982). The outfit statistic is simply an unweighted average of these squared standardized residuals (e.g., Equation 3).

$$E_{ni} = \sum_{x=0}^{m} xP(X_{ni} = x), x = 0, 1, \ldots, m \quad (2)$$

where E_{ni} is the expected value for a combination of person n and item i which has $m + 1$ rating scale categories. $P(X_{ni} = x)$ is given in Equation 1.

$$outfit_{\hat{\beta}} = \frac{\sum_{i=1}^{I} z_{ni}^2}{I} \quad (3)$$

where $z_{ni} = \frac{x_{ni} - E_{ni}}{\sqrt{V(x_{ni})}}$

and $V(x_{ni}) = \sum_{x=0}^{m}(j - E_{ni})P(X_{ni} = x), x = 0, 1, ..., m$

Outfit statistics are reported as a chi-square statistic, divided by its degrees of freedom which results in an expected value of 1.00 and a range from 0.00 to ∞ (Linacre and Wright, 1994). A mean square outfit statistic greater than 1.00 suggests the presence of unexpected residuals in the tails of the score distribution, and a mean square outfit value less than 1.00 indicates less variability than expected based on the RSM. In general, elements with mean square outfit statistics ranging from 0.8 to 1.4 are considered to show adequate fit to the model (Wright and Linacre, 1994), although the cutoff values tend to vary depending on the purpose for which the ratings are used.

Mean square outfit statistics can be standardized (z_{outfit}) to have a mean of 0.00 and standard deviation of 1.00 (e.g., Equation 4) (Wright and Masters, 1982). Estimates with $| z | > 2.00$ exhibit poor fit to the model and should be further examined to determine whether there are problems with the scores associated with that particular examinee, item, or category threshold.

$$z_{outfit} = \left(\sqrt[3]{outfit} - 1\right)\left(\frac{3}{V(outfit)}\right) + \left(\frac{V(outfit)}{3}\right) \quad (4)$$

An important feature of the RSM is that it provides a means for evaluating the extent to which person measures are stable across samples of items and item calibrations are stable across samples of people (i.e., **parameter invariance**). In the present context, invariance evaluation provides a means for determining the extent to which measures for individual persons or calibrations for individual items exhibit changes across measurement occasions. The stability of two parameter estimates ($\hat{\theta}_1$ and $\hat{\theta}_2$) that are obtained on different occasions is evaluated by examining the standardized difference (Equation 5) between the two estimates (Wright and Masters, 1982). The standardized differences for a population or an item pool that conforms to the RSM have an expected value of 0.00 and an expected standard deviation of 1.00. Large departures in observed data from these expected values indicate estimates that are more or less stable over time than would be expected.

$$z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{[SE(\hat{\theta}_1)]^2 + [SE(\hat{\theta}_2)]^2}} \quad (5)$$

Problem

Data

Data from this study were taken from an evaluation of a family care program in which parents, primarily mothers, received interventions throughout a year-long program. The foci of the program codified in the curriculum were parent-child interactions, family functioning, parent psychological functioning, parents' use of community services, and parents' and children's use of medical services (Stoffelmayr, Reischl, Lounsbury, and Chiu, 1996). Participants were interviewed on three occasions. They received intervention including counseling, workshops, parent education, and social support throughout the intervention year. The scale measuring parent psychological functioning was analyzed for this study.

Measures for Occasion 1 were taken immediately prior to participation in the program. Measures for Occasion 2 were taken four months after the initiation of the intervention. Measures at Occasion 3, serving as a maintenance check, were taken six months after the intervention. Hence, if the intervention was successful, we would expect each person measure to increase (improved psychological functioning) between Occasions 1 and 2 and to remain at about the same level (possibly with a slight regression toward the pre-intervention levels) between Occasions 2 and 3.

Instrument

Altogether, 128 program participants responded to a 14-item self-report rating scale containing 4 Likert-type options. The 14 items were a subset of the General Health Quesionnaire (Goldberg, 1978) that were selected to assess general health status including anxiety and depression levels (Stoffelmayr, Reischl, Lounsbury, and Chiu, 1996). The construct underlying this questionnaire, then, indicates the frequency that program participants exhibited anxious and depressed behaviors with higher levels of the scale indicating **fewer** episodes of these behaviors and lower levels of the scale indicating **more** episodes of these behaviors.

Preliminary Analyses

These data were analyzed by applying the RSM to data from each occasion separately. As a result, we obtained three measures for each person, three calibrations for each item, and three category thresholds (one estimate for each occasion). In these analyses, people were positively scaled so that positive logits were associated with higher raw score values, indicating a healthier psychological status (i.e., less anxious and depressed) on the underlying continuum. Items were negatively scaled so that positive logits were associated with lower raw scores, indicating an item that was more difficult to endorse. The average item difficulty in each calibration was centered. Parameter estimation was performed using *Facets* (Linacre, 1989), although these analyses could be performed using most Rasch analysis programs.

As previously mentioned, the intervention was designed to decrease a participant's numbers of anxiety and depression episodes. This intervention was imposed between the first and second occasions and was removed between the second and third occasions. Hence, we expected to observe an increase in the person measures from Occasion 1 to Occasion 2 and a slight diminishing of that effect between Occasions 2 and 3. In addition, because the intervention could potentially change the way that the person feels about his/her mental health, we expected to see a slight redefinition of the variable being measured as a result of the intervention. That is, we expected to observe slight changes in the ordering of item calibrations across time. Unfortunately, our expectations were not supported by our preliminary results.

Preliminary Results

Interpretation of the preliminary analyses focused on three features: a) evaluation of the rating scale structure, b) interpretation of the item calibrations, and c) interpretation of the person measures.

Rating Scale Calibrations

To create meaningful measures using Likert-type questionnaire items across multiple occasions, several criteria need to be met (Linacre, 1999). First, there should be a minimum of ten observations in each category. Second, the shape of the rating scale distribution should be peaked. Third, the average category measures should increase with the rating scale categories. Fourth, the outfit mean square statistics should lie within a rea-

sonable range of values. Fifth, category thresholds should increase with the rating scale categories. Sixth, the category thresholds should be at least 1.4 logits apart and more than 5 logits apart. In addition, the category threshold calibrations need to be invariant across the occasions of measurement (Smith, 1997; Wilson, 1992; Wright, 1996b).

Evidence of the validity of our rating scale is provided in Table 1, which reveals that there are problems with the rating scale. The rating scale meets several of the requirements identified above. For example, each rating scale category is used with adequate frequency, and the distribution of responses are unimodal. Also, the average measures and category threshold calibrations increase with the rating scale categories. However, an inspection of the mean square outfit statistics reveals that rating scale Category 2 on Occasion 2, Category 3 on Occasion 3, and Category 4 on all 3 occasions show poor fit to the RSM.[1] In addition, the rating scale estimates are not invariant across occasions. The category threshold calibrations for the first threshold are more variable than expected across all three occasions, and the threshold calibrations for the second threshold are more variable than expected between Occasions 1 and 3.[2]

Item Calibrations

Table 2 summarizes the results of our preliminary analyses of the uncorrected item calibrations. Examination of this table reveals several additional problems with the measures from our evaluative design. Several of the items show poor fit to the RSM—two on Occasion 1 (Items 8 and 12), three on Occasion 2 (Items 3, 9, and 11), and three on Occasion 3 (Items 4, 6, and 13). This is reiterated by the fact that the standard deviation of the standardized mean square outfit statistics is greater than the expected value of 1.00. In addition, several of the item parameters (21% between Occasions 1 and 2 and 36% between Occasions 1 and 3) demonstrate parameter instability over time. In fact, only six of the fourteen items (43%) **do not** exhibit parameter instability at either occasion transition. The item separation reliabilities for the three occasions were .97 on all three occasions, indicating that these distributions of item parameters contain enough variability to create distinct strata of item difficulties. Item separation indices for the three occasions were 5.85, 6.06, and 6.03.

Table 1

Uncorrected Rating Scale Calibrations, Standard Errors, Fit Statistics, and Standardized Differences

Category	Percent (Count)			Average Measure			Category Threshold (Standard Error)			Mean Square Outfit[a]			Standardized Differences[b]	
	O1	O2	O3	O1	O2	O3	O1	O2	O3	O1	O2	O3	O1-2	O1-3
1	67 (1183)	64 (1133)	63 (1118)	-2.88	-2.91	-3.48	— —	— —	— —	1.1	1.0	1.0	—	—
2	20 (354)	22 (396)	488 (28)	-1.21	-1.13	-1.41	-.81 (.06)	-1.02 (.06)	-1.49 (.06)	.8	*.7*	1.2	*2.47*	*8.01*
3	9 (161)	10 (175)	8 (136)	-.41	-.23	-.68	.00 (.09)	-.01 (.09)	.35 (.09)	1.2	.8	*2.4*	0.08	*-2.75*
4	4 (77)	4 (70)	2 (31)	.48	-.05	*-.22*	.81 (.14)	1.03 (.14)	1.14 (.19)	*1.6*	*2.1*	*1.7*	-1.11	-1.40

[a] Mean square outfit indices less than .8 and greater tha 1.4 are considered to indicate rating scale misfit.

[b] Absolute standardized differences greater than 2.00 are considered large enough to indicate rating scale invariance across time.

Table 2

Uncorrected Item Calibrations, Standard Errors, Fit Statistics, and Standardized Differences

Item	Item Calibration			Standard Error			Standardized Mean Square Outfit[a]			Standardized Differences[b]	
	O1	O2	O3	O1	O2	O3	O1	O2	O3	O1-2	O1-3
1	-1.21	-1.17	-1.67	0.12	0.12	0.12	0.29	0.99	-0.71	-0.24	*2.71*
2	1.11	1.62	2.30	0.25	0.28	0.46	1.65	1.27	0.05	-1.36	*-2.27*
3	-0.34	-0.13	-0.40	0.15	0.15	0.16	1.01	*2.32*	1.53	-0.99	0.27
4	-1.65	-1.84	-2.35	0.11	0.11	0.11	0.03	-0.10	*-3.43*	1.22	*4.50*
5	1.11	1.48	1.95	0.25	0.26	0.39	*3.61*	1.31	1.50	-1.03	-1.81
6	-1.44	-1.09	-1.72	0.11	0.12	0.12	*-2.40*	-1.34	*-2.27*	*-2.15*	1.72
7	0.99	0.47	0.80	0.23	0.18	0.24	0.51	0.44	1.22	1.78	0.57
8	1.30	1.23	1.68	0.27	0.24	0.34	1.67	0.97	1.11	0.19	-0.88
9	-1.28	-0.76	-0.99	0.12	0.13	0.14	-0.14	*2.58*	1.40	*-2.94*	-1.57
10	2.45	1.28	1.68	0.45	0.24	0.34	0.44	1.91	0.72	*2.29*	1.37
11	-1.06	-1.11	-1.43	0.12	0.12	0.13	-1.70	*-3.83*	-1.56	0.29	*2.09*
12	-0.78	-0.87	-1.46	0.13	0.13	0.13	-1.83	-1.21	-1.78	0.49	3.70
13	-0.50	-0.33	-0.69	0.14	0.14	0.15	0.55	1.29	*2.94*	-0.86	0.93
14	1.30	1.22	2.29	0.27	0.24	0.46	0.19	0.88	*2.07*	0.22	-1.86
Mean	0.00	0.00	0.00	0.19	0.18	0.24	0.28	0.53	0.20	-0.22	0.68
(SD)	(1.32)	(1.19)	(1.70)	(0.10)	(0.06)	(0.13)	(*1.55*)	(*1.70*)	(*1.86*)	(1.46)	(2.16)

[a] Absolute standardized mean square fit indices greater than 2.00 are considered large enough to indicate item misfit.

[b] Absolute standardized differences greater than 2.00 are considered large enough to indicate item invariance across time.

Person Measures

Table 3 summarizes the uncorrected person measures. This table shows that the average person measure remained about the same between Occasions 1 and 2 but that the measures decreased slightly at Occasion 3. About 20% of the people measures exhibited statistical differences between Occasions 1 and 2, and about 28% of the measures were statistically different between Occasions 1 and 3. In addition, a fairly large proportion of the people did not fit the RSM, particularly on Occasions 2 and 3 (supported by the inflated standard deviation of the standardized mean square outfit statistic). Person separation reliability was somewhat low on Occasions 1 and 2 (.72 and .75, respectively), and the reliability index was very low on Occasion 3 (.46). Separation indices for the three occasions were 1.61, 1.74, and 0.92.

Table 3

Summary Statistics for Uncorrected Person Measures, Fit Statistics, and Standardized Differences

	Person Logits			Standard Errors			Standardized Mean Square Outfit[a]			Standardized Differences		Percent Exhibiting Change	
	O1	O2	O3	O1	O2	O3	O1	O2	O3	O1-2	O1-3	O1-2	O1-3
Mean	-2.17	-2.14	-2.61	0.52	0.52	0.52	-0.26	-0.24	-0.36	-0.06	0.71	*20%*	*28%*
SD	1.04	1.11	0.72	0.18	0.18	0.10	*1.23*	*1.47*	*1.33*	1.61	1.64	—	—

[a]10% of the people misfit the RSM on Occasion 1, 16% on Occasion 2, and 20% on Occasion 3.
Note: $N = 128$

Preliminary Interpretations

The preliminary rating scale and item analyses indicate several potential problems with the measurement of psychological functioning across the three occasions. Some of the category thresholds exhibit poor fit, and the category threshold calibrations are not all invariant over time. In addition, several of the items exhibit poor fit to the RSM, and more than half of the item parameter estimates are unstable across occasions. As a result, we have a situation in which the facets of our measurement device that are used to establish the continuum for the underlying latent trait are too variable to allow for measurement of changes in people over time. That is, the fact that the category thresholds were used differently and the item orderings were not consistent across occasions implies that slightly different constructs are being measured at each time point.

Perhaps these problems explain the confusing results of the preliminary person analyses. Perhaps the changes in rating scales and items across

occasions introduces ambiguity into our interpretations of changes in people. The results that we have obtained seem completely contradictory to what we would expect, and we would have a difficult time explaining how these results were obtained. We had expected that the intervention would increase person psychological functioning between Occasions 1 and 2 and that person psychological functioning would (hopefully) be maintained at the 6-month follow-up (between Occasions 2 and 3). However, our preliminary results suggest that there was no average change in persons between the pre-intervention and post-intervention and that measures **decreased** (to a level lower than pre-intervention) at the follow-up.

Method

As previously mentioned, Wright (1996b) proposed an algorithm for disentangling changes in measures of multiple facets of the measurement context. Previously, Wolfe and Chiu (1999) demonstrated how this algorithm can be applied to data from an evaluative study that is taken from two occasions, and a detailed description of the algorithm (including *Facets* code) is available from that article. Here we summarize the algorithm, highlighting how it can be extended to disentangle changes in measures of multiple facets across three or more occasions. An overview of the algorithm is shown in Figure 1.

Step 1—Evaluate Rating Scale and Item Invariance

The first step of the procedure is to evaluate rating scale and item invariance across the measurement occasions (as we did in the preliminary analyses). That is, data from each time point are calibrated to the RSM independently (i.e., in separate analyses). Parameter estimates are compared using the standardized difference (Equation 5). If all category threshold or item parameter estimates exhibit invariance across occasions, then no correction is needed. If category thresholds or item calibrations are unstable across occasions, however, then the remaining steps of the algorithm are needed. In the preliminary analyses, we found that the first category threshold is unstable across both occasions, and the second category threshold is unstable between Occasions 1 and 3 (refer to Table 1). In addition, we found that seven of the fourteen items exhibit instability on at least one of the occasion transitions (refer to Table 2). Therefore, we would proceed with the remaining four steps of the algorithm.

STEP 1

Evaluate rating scale and item invariance with separate analyses data from each occasion. Note: rating scale invariance is evaluated via τ_{x1}-τ_{x2} and τ_{x1}-τ_{x3}, and item invariance is evaluated via δ_{i1} - δ_{i2} and δ_{i1}-δ_{i3}.

↓

STEP 2

Create common category threshold calibrations (τ_{xc}) by stacking data from all three occasions, treating persons as unique at each time point.

↓

STEP 3

Create corrected person measures (β_{n1c}) and item calibrations (δ_{i1c}) for Occasion 1 by anchoring the rating scale on τ_{xc}.

↓

STEP 4

Create corrected person measures for the remaining occasions (β_{n2c} and β_{n3c}) by anchoring the rating scale on τ_{xc} and anchoring the items on δ_{j1c} (only the *J* items that are invariant across all occasions).
Note: person measure invariance is evaluated via β_{n1c}-β_{n2c} and β_{n1c}-β_{n3c}.

↓

STEP 5

Create corrected item calibrations for the remaining occasions (δ_{i2c} and δ_{i3c}) by anchoring the rating scale on τ_{xc} and anchoring the people on β_{n1c}.
Note: item calibration invariance is evaluated via δ_{i1c}-δ_{i2c} and δ_{i1c}-δ_{i3c}.

Figure 1. Steps for creating a frame of references for interpreting change across three measurement occasions.

Step 2—Create a Common Rating Scale

The second step in disentangling changes in rating scale and item calibrations from changes in person measures is to create a stable frame of reference for the measurement context. We do this by creating a common rating scale structure to which the remaining facets of our measurement context can be anchored. We accomplish this by stacking data from all three

time points in a single data set, maintaining the item identity across the three occasions but treating each person as unique at each time period. For our example, we added 1,000 to each person identifier for the Occasion 2 data and added 2,000 to each person identifier for the Occasion 3 data. We calibrated this data set to the RSM. The resulting category threshold calibrations (τ_{xc}) constitute an "average" underlying rating scale that can be used to describe the data at all three occasions. We use these calibrations as rating scale anchor values in the remaining steps of the algorithm.

Step 3—Correct Occasion 1 Estimates

The third step in the correction algorithm is to correct the person measures and item calibrations at Occasion 1 by anchoring them to the underlying rating scale that was created in Step 2. To accomplish this, we calibrate the Occasion 1 data to the RSM using the Step 2 category threshold calibrations (τ_{xc}) as anchor values, person and item parameters were free to vary. The resulting corrected person measures (β_{n1c}) and item calibrations (δ_{i1c}) are used in subsequent steps as benchmarks against which the remaining occasions are compared.

Step 4—Correct Subsequent Occasion Person Measures

Once corrected item calibrations have been obtained for Occasion 1 (in Step 3), these calibrations are used to anchor the invariant items so that person measures that are free from the influence of changing items can be obtained for the remaining occasions. That is, the corrected item calibration values for Occasion 1 (from Step 3) are used to create a common instrument across all three occasions so that person measures can be interpreted against a stable continuum. To accomplish this, calibrate data from each of the remaining occasions to the RSM while anchoring items and anchoring category thresholds on their corrected values (δ_{j1c} and τ_{xc}, respectively). It should be noted that **only the J items that were found to be invariant across all three measurement occasions are anchored**—the remaining (I - J) items are allowed to float (i.e., they are treated as different items on subsequent occasions). In our example, this step results in corrected person measures for Occasion 2 (β_{n2c}) and Occasion 3 (β_{n3c}). Change in persons over time is interpreted by comparing the corrected measures for Occasions 2 and 3 to the corrected measures for Occasion 1 (using the standardized difference).

Step 5—Correct Subsequent Occasion Item Calibrations

In the final step of the algorithm, we correct item calibrations for Occasions 2 and 3 by anchoring person measures in each data set to their corrected values (obtained in Step 4). That is, we calibrate Occasion 2 data to the RSM by anchoring persons on their corrected Occasion 2 measures (β_{n2c}) and anchoring category thresholds on their corrected calibrations (τ_{xc}) allowing all item parameters to vary. This results in corrected Occasion 2 item calibrations (δ_{i2c}) for all items. Similarly, we calibrate Occasion 3 data to the RSM, anchoring persons on β_{n3c} and category thresholds on τ_{xc}, to obtain corrected Occasion 3 item calibrations (δ_{i3c}). Change in items over time is interpreted by comparing the corrected calibrations for Occasions 2 and 3 to the corrected calibrations for Occasion 1 (using the standardized difference).

Results

Interpretation of the analyses of the corrected data focused on four features: a) evaluation of the rating scale structure, b), interpretation of the person measures, c) interpretation of the item calibrations, and d) methodological comparisons.

Rating Scale Calibrations

The corrected rating scale calibrations, obtained from Step 2, are summarized in Table 4. This table shows that the corrected rating scale calibrations seem to describe the data across the three occasions reasonably well. The percentage of observations falling into each rating scale

Table 4

Corrected Rating Scale Calibrations, Standard Errors, and Fit Statistics

Category	Percent (Count)	Average Measure	Category Thresholds (Standard Error)	Mean Square Outfit[a]
1	65 (3434)	-3.03	— —	1.0
2	23 (1238)	-1.22	-1.10 (0.04)	0.8
3	9 (472)	-0.43	0.11 (0.05)	1.3
4	3 (178)	0.16	0.99 (0.09)	*1.8*

[a]Mean square outfit indices less than .8 and greater than 1.4 are considered to indicate rating scale misfit.

category seem reasonable, the average measures and category threshold calibrations increase with the rating scale values, and three of the four ratings scale categories show good fit to the RSM. The fourth (highest) rating scale category shows slight misfit.

Person Measures

Table 5 summarizes the corrected person measures, which were obtained in Steps 3 and 4. As shown, the corrected measures increase between Occasion 1 and Occasion 2 and decrease to near their Occasion 1 values at Occasion 3—something that we would predict, given that the intervention was administered between Occasions 1 and 2 and was removed between Occasions 2 and 3. Also, note that about 9% of the people misfit the RSM on Occasion 1, but a larger proportion of the people misfit on Occasions 2 and 3—again, something we would expect if the instrument was designed to indicate pre-treatment needs. In addition, note that

Table 5

Summary Statistics for Corrected Person Measures, Fit Statistics, and Standardized Differences

	Person Logits			Standard Errors			Standardized Mean Square Outfit[a]			Standardized Differences		Percent Exhibiting Change	
	O1	O2	O3	O1	O2	O3	O1	O2	O3	O1-2	O1-3	O1-2	O1-3
Mean	-2.37	-1.86	-2.21	0.54	0.36	0.31	-0.09	-0.06	-0.48	-0.69	0.04	31%	32%
SD	1.14	1.15	0.64	0.17	0.14	0.09	1.27	1.48	1.31	1.94	2.06	—	—

[a] 9% of the people misfit the RSM on Occasion 1, 16% on Occasion 2, and 15% on Occasion 3.
Note: $N = 128$

about one-third of the people measures showed instability at the Occasion 1 and 2 and the Occasion 1 and 3 transitions. The person measure reliabilities were .75, .88, and .74 at Occasions 1, 2, and 3 (respectively), and the corresponding person separation indices were 1.73, 2.77, and 1.68.

Item Calibrations

The corrected item calibrations, obtained from Steps 3 and 5, are summarized in Table 6. This table shows that there are several problems with the items. For example, ten of the fourteen items show poor fit to the RSM on at least one of the occasions. In addition, eleven of the fourteen items show parameter instability across at least one of the occasion transitions. As a result, we cannot be sure that the construct being measured

Table 6

Corrected Item Calibrations, Standard Errors, Fit Statistics, and Standardized Differences

Item	Item Calibration			Standard Error			Standardized Mean Square Outfit[a]			Standardized Differences[b]	
	O1	O2	O3	O1	O2	O3	O1	O2	O3	O1-2	O1-3
1	-1.32	-0.89	-1.66	0.13	0.11	0.10	1.14	1.24	*-2.27*	*-2.53*	*2.07*
2	1.21	1.11	2.28	0.25	0.13	0.34	*2.04*	-1.05	0.00	0.35	*-2.54*
3	-0.36	0.19	-0.56	0.15	0.12	0.13	1.68	*2.54*	-0.39	*-2.86*	1.01
4	-1.81	-1.58	-1.93	0.12	0.11	0.08	0.90	0.14	*-4.69*	-1.41	0.83
5	1.21	0.95	1.94	0.25	0.12	0.19	*4.09*	-1.31	1.34	0.94	*-2.32*
6	-1.57	-0.81	-1.70	0.12	0.12	0.09	-1.59	-1.11	*-4.01*	*-4.48*	0.87
7	1.09	0.80	0.84	0.24	0.16	0.20	0.79	0.61	0.81	1.01	0.80
8	1.42	0.69	1.68	0.28	0.13	0.29	*2.02*	-1.21	0.95	*2.36*	-0.64
9	-1.39	-0.46	-1.08	0.12	0.11	0.11	0.70	*2.81*	-0.60	*-5.71*	-1.90
10	2.63	0.75	1.68	0.46	0.13	0.29	0.53	-0.57	0.57	*3.93*	1.75
11	-1.15	-0.83	-1.45	0.13	0.10	0.09	-0.94	*-3.59*	*-3.34*	-1.95	1.90
12	-0.84	-0.58	-1.48	0.14	0.12	0.09	-1.14	-0.97	*-3.10*	-1.41	*3.85*
13	-0.53	-0.02	-0.82	0.15	0.13	0.12	1.23	1.49	0.88	*-2.57*	1.51
14	1.42	0.68	2.27	0.28	0.12	0.31	0.42	-1.57	1.95	*2.43*	*-2.03*
Mean	0.00	0.00	0.00	0.20	0.12	0.17	0.85	-0.18	-0.85	-0.85	0.37
(SD)	1.44	0.85	1.67	0.10	0.01	0.10	1.46	1.76	2.20	2.78	1.95

[a] Absolute standardized mean square outfit indices greater than 2.00 are considered large enough to indicate item misfit.
[b] Absolute standardized differences greater than 2.00 are considered large enough to indicate item invariance across time.

at Occasion 1 is the same as the construct being measured at Occasion 2, and/or Occasion 3. The corrected item separation reliabilities for the three occasions were .97, .98, and .98 (for Occasions 1, 2, and 3, respectively). The corresponding item separation indices were 6.16, 6.65, and 8.09.

Methodological Comparisons

A comparison of the conclusions that one would draw based on the corrected and the uncorrected item calibrations and person measures reveals some substantial differences between the two methods of analysis. With respect to the rating scale calibrations (τ_x), there are only minor differences between the uncorrected and corrected calibrations. The uncorrected calibrations exhibit parameter invariance across occasions and exhibit misfit at two of the three occasions. The values of the corrected calibrations are close to the average value of the three uncorrected calibrations. In addition, the corrected values show slightly better fit, on average, to the RSM.

The differences between Tables 2 and 4 (the uncorrected and corrected item calibration summaries) are shown in Table 7. As shown in this table, the standardized differences that are based on the corrected item calibrations suggest that, overall, there was more change in the items between Occasions 1 and 2 than between Occasions 1 and 3. This makes sense. The manifestation of the underlying construct should change as a result of the intervention, and we would expect the definition of that construct to begin to resemble the pre-intevention definition after removal of the intervention. And, as would be expected, a larger proportion of standardized differences were statistically significant with the corrected values than with the uncorrected values. If we were to compare the corrected and uncorrected standardized differences, our decisions concerning item invariance would change for about one-third of the items as a result of

Table 7

Comparison of Uncorrected and Corrected Item Calibrations

	Standardized Difference[a]				Standardized Mean Square Outfit					
	Uncorrected		Corrected		Uncorrected			Corrected		
	O1-2	O1-3	O1-2	O1-3	O1	O2	O3	O1	O2	O3
Percent Significant	21	36	36	57	14	21	29	21	21	36
Mean	-0.22	0.68	-0.85	0.37	0.28	0.53	0.20	0.85	-0.18	-0.85
SD	1.46	2.16	2.78	1.95	1.55	1.70	1.86	1.46	1.76	2.20

[a]36% of the decisions about individual items would change between Occasions 1 and 2 if corrected calibrations were considred instead of uncorrected calibrations, and 29% of the decisions about individual items would change between Occasions 1 and 3.

Note: $N = 14$

this correction procedure. In addition, the items show poorer fit to the RSM using the corrected values.

The benefit of using this correction procedure is most apparent when one examines the differences between the corrected and uncorrected person measures. These differences are shown in Table 8 (a summary of the differences between Tables 3 and 5). First, notice that the uncorrected person measures show no change between Occasions 1 and 2 and a sharp drop between Occasions 1 and 3. On the other hand, the corrected measures show an increase between Occasions 1 and 2 and a decrease between Occasions 1 and 3. The latter of these makes more sense than the former. This trend is shown more clearly in Figure 2, a plot of the average corrected and uncorrected person measures across the three occasions.

Table 8

Comparison of Uncorrected and Corrected Person Measures

	Standardized Difference[a]				Standardized Mean Square Outfit					
	Uncorrected		Corrected		Uncorrected			Corrected		
	O1-2	O1-3	O1-2	O1-3	O1	O2	O3	O1	O2	O3
Percent Significant	20	28	31	32	10	16	20	9	16	15
Mean	-0.06	0.71	-0.69	0.04	-0.26	-0.24	0.36	-0.09	-0.06	-0.48
SD	1.61	1.64	1.94	2.06	1.23	1.47	1.33	1.27	1.48	1.31

[a] 15% of the decisions about individual items would change between Occasions 1 and 2 if corrected measures were considred instead of uncorrected measures, and 23% of the decisions would change between Occasions 1 and 3.

Note: $N = 128$

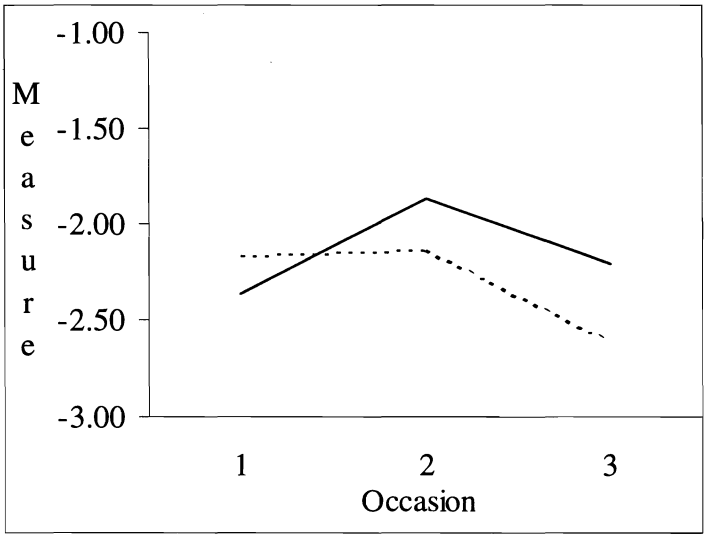


Figure 2. Average corrected and uncorrected person measures across the three measurement occasions.

Here, it is apparent that the corrected measures (solid line) show the expected rise (between pre-intervention and post-intervention) and fall (after post-intervention) while the uncorrected measures (dotted line) portray a pattern that is difficult to explain. Additional encouraging information is shown in Table 8 by the fact that a larger proportion of individual people exhibit change with the corrected measures than with the uncorrected measures. And, as would be expected, a substantial number of people (about 20% of the total *N*) would have different decisions made about the significance of their changes if decisions were based on corrected versus uncorrected measures. Finally, note that people tend to show a better fit to the RSM once the influence of unstable rating scale and item calibrations are removed from their measures.

Discussion

In this article, we have demonstrated the practical utility of an algorithm for disentangling changes in different facets of a measurement context. If we had simply interpreted the uncorrected data, we would have had a difficult time explaining the confusing pattern of person measures across time. By removing the instability of rating scale calibrations and basing person measures on a stable core of items, we were able to identify a meaningful trend in the person measures. In addition, we found that the correction procedure increased the number of people who were suspected of changing as a result of the intervention and that the people, overall, showed better fit to the RSM.

The primary limitation of this study is that we cannot be positive that the corrected measures are actually better measures than the uncorrected measures. We certainly suspect that this is true, and we believe that the evidence provided here makes a strong argument for the validity of the corrected measures. However, without controlled simulation studies to determine the power with which the correction method recovers true changes in item and person parameters, we can only speculate that the correction procedure is actually accomplishing the goal for which it was intended. Hence, one direction for future research concerning this correction algorithm would be to perform simulations to determine the efficiency and effectiveness of this algorithm for disentangling changes in rating scales and items from changes in persons across time.

In addition, there are several similar methods that might be employed to disentangle the changes that are of concern here. For example,

Wright (1996a) suggests three ways of establishing stability within a Rasch measurement context, Chang and Chan (1995) compared several methods for establishing this stability in an applied setting. What is needed at this point in time is a controlled comparison of each of these potential alternatives, again under simulated conditions, to determine which ones provide the most accurate portrayal of true changes in the multiple facets of the measurement situation.

Finally, we suggest that another interesting direction to further our understandings of these procedures would be to extend these correction algorithms to multi-faceted situations. Wolfe and Chiu (1999) have demonstrated how the Wright (1996b) algorithm can be employed to measure change across two measurement occasions. And, in the current article, we have shown how that algorithm can be extended to measure change across three or more occasions. Many instruments used in program evaluations require multi-faceted measurement models. For example, many rehabilitation and educational evaluations employ raters as a component of the measurement procedures. Extending algorithms for disentangling changes within a measurement context to such multi-faceted designs would be another important step in evaluating the utility of these methods.

Footnotes

[1] We considered mean square outfit statistic values ranging from 0.8 to 1.4 to be adequate.

[2] We chose to depict all change from the pre-intervention baseline measure. That is, we evaluated invariance by comparing Occasion 2 measures to Occasion 1 measures and by comparing Occasion 3 measures to Occasion 1 measures.

References

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

Chang, W. C and Chan, C. (1995). Rasch analysis for outcome measures: Some methodological considerations. *Archives of Physical Medicine and Rehabilitation, 76*, 934-939.

Cook, T. C. and Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston, MA: Houghton Mifflin.

Goldberg, D. P. and Hiller, V. F. (1979). A scaled version of the General Health Questionnaire. *Psychological Medicine, 9*, 139-145.

Linacre, J. M. (1989). *Many-facet Rasch measurement.* Chicago, IL: MESA Press.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3,* 103-122.

Roderick, M. and Stone, S. (1996, April). Is it changing opinions or changing kids? Paper presented at the 1996 annual meeting of the American Educational Research Association, New York, NY.

Smith, R. M. (1997). Pre/post comparisons in Rasch measurement. In M. Wilson, K. Draney, and G.J. Engelhard (eds.), *Objective measurement: Theory into practice* (Vol. 4) (pp. 297-312). Greenwich, CT: Ablex.

Stoffelmayr, B., Reischl, T., Lounsbur, D., and Chiu, C. W. T. (1996). *Detroit Skillman parenting program outcome evaluation final report.* Department of Psychiatry, Michigan State University: E. Lansing, Michigan.

Wilson, M. (1992). Measuring changes in the quality of school life. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2) (pp. 77-96). Greenwich, CT: Ablex.

Wolfe, E. W. and Chiu, C. W. T. (1999). Measuring pretest-posttest change with a Rasch rating scale model. *Journal of Outcome Measurement, 3,* 134-161.

Wright, B. D. (1996a). Comparisons require stability. *Rasch Measurement Transactions, 10,* 506.

Wright, B. D. (1996b). Time 1 to time 2 comparison. *Rasch Measurement Transactions, 10,* 478-479.

Wright, B. D. and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8,* 370.

Wright, B. D. and Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.

JOURNAL OF OUTCOME MEASUREMENT®, 3(4), 382–405
Copyright© 1999, Rehabilitation Foundation, Inc.

Understanding Rasch Measurement: Estimation Methods for Rasch Measures

John M. Linacre
University of Chicago

Rasch parameter estimation methods can be classified as non-interative and iterative. Non-iterative methods include the normal approximation algorithm (PROX) for complete dichotomous data. Iterative methods fall into 3 types. Datum-by-datum methods include Gaussian least-squares, minimum chi-square, and the pairwise (PAIR) method. Marginal methods without distributional assumptions include conditional maximum-likelihood estimation (CMLE), joint maximum-likelihood estimation (JMLE) and log-linear approaches. Marginal methods with distributional assumptions include marginal maximum-likelihood estimation (MMLE) and the normal approximation algorighm (PROX) for missing data. Estimates from all methods are characterized by standard errors and quality-control fit statistics. Standard errors can be local (defined relative to the measure of a particular item) or general (defined ralative to the abstract origin of the scale). They can also be ideal (as though the data fit the model) or inflated by the misfit to the model present in the data. Five computer programs, implementing different estimation methods, produce statistically equivalent estimates. Nevertheless, comparing estimates from different programs requires care.

Requests for reprints should be sent to John M. Linacre, MESA Psychometric Laboratory, University of Chicago, 5835 S. Kimbark Avenue, Chicago, IL 60637, e-mail: mesa@uchicago.edu.

Rasch measurement is the only way to convert ordinal observations into linear measures (Fischer, 1995). These measures are represented as parameters in a Rasch model and are estimated from ordinal data. The analyst, however, is rarely concerned about the estimation process, provided that reasonable values for the measures are obtained. The precision of the measures can be characterized by their standard errors, and their statistical validity by fit statistics. An appreciation of the different methods of estimation, however, enables the analyst to better evaluate what constitute reasonable measures.

When a new measurement situation is encountered, the Rasch measures, the parameters of a relevant Rasch model, must be inferred from data. This is accomplished by means of the method of inverse probability, first described by Jacob Bernoulli (1713). Inverse probability enables us to estimate values for the measures, but those values are always approximate to some degree. The fact that all measures (including Rasch measures) are approximate, is rarely of major concern, because "for problem solving purposes, we do not require an exact, but only an approximate, resemblance between theoretical results and experimental ones" (Laudan, 1977).

Estimation, Precision and Accuracy

Rasch estimates are always characterized by their precision and their accuracy. In this context, precision relates to the uncertainty in the measure, the estimated location of the parameter on the latent variable, when it is specified that the data fit the Rasch model. This precision is reported as the standard error of the measure. When the data are specified to fit a Rasch model, then all unexpectedness in the data are deemed to be products of the probabilistic processes required by Rasch models.

Precision can always be increased by collecting more relevant data or specifying rating scales with more categories, with the continuing condition that the data are specified to fit the model. Precision can be artificially improved by introducing constraints, often as assumptions, which reduce the location uncertainty. The most commonly introduced assumption is that one or more characteristics underlying the data are normally distributed.

Accuracy relates to the departure of the data from those values predicted by a Rasch model given the estimated locations of the parameters. The degree of departure is summarized in fit statistics and other indica-

tors of conformity of the data to the Rasch model. No empirical data set fits the Rasch model perfectly. Nevertheless, as the data depart ever further from meeting Rasch model expectations, doubt not only about the locations, but also about the meaning of parameter estimates increases. Accuracy can be increased by collecting more data that is likely to conform to a Rasch model, e.g., by avoiding administering items that are too trivial or too challenging, which are likely to provoke irrelevant behavior in respondents. Accuracy can also be increased by screening out responses deemed irrelevant for measurement purposes. Such responses may be highly diagnostic of idiosyncratic aspects of respondents, items, judges or the rating scale, but they do not contribute to constructing a generalizable measurement system.

Due to the arbitrary nature of pass-fail decisions and the practical need to introduce determinacy into both norm-referenced and criterion-referenced reporting, Rasch estimates (as well as raw scores and other statistics) are usually treated as point-estimates of their underlying parameters. Thus estimates are commonly reported with more significant figures than either their precision or their accuracy supports. Since all estimation methods are approximate, the same estimation method under different conditions, or different estimation methods under the same conditions, may disagree numerically as to whether a subject, near to the pass-fail point, is a "pass" or a "fail".

The Nature of the Rasch model

Consider the basic Rasch model. This postulates that the data are the dichotomous outcomes of a probabilistic process governed by a linear combination of parameters, called here the person ability and the item difficulty. All estimation methods in common use for other Rasch models can also be applied to the basic model. This model is:

$$\log\left(\frac{P_{ni1}}{P_{ni0}}\right) \equiv B_n - D_i, \qquad (1)$$

where

B_n is the ability of subject n, where $n = 1,N$,
D_i is the difficulty of item i, where $i = 1,L$,
P_{ni1} is the probability that subject n will succeed on item i,
P_{ni0} is the probability of failure $1-P_{ni1}$.

P_{ni1} can be expressed as:

$$P_{ni1} = \frac{e^{(B_n - D_i)}}{1+e^{(B_n - D_i)}}. \quad (2)$$

Were the model parameters to be known, then the probability of observing any particular datum would also be known. Each data point, X_{ni}, has a value of x, which is 1 if person n succeeds on item i, and 0 otherwise. The expected value of the datum, E_{ni}, is P_{ni1}. For any parameter, e.g., n (or i), the marginal score, R_n (or R_i), the sum of all observations modeled to be generated by n (or i), is

$$R_n = \sum_n X_{ni} \approx \sum_n E_{ni} = \sum_n P_{ni1}. \quad (3)$$

Thus the frequency of successful responses in the data becomes the basis for inferring the probabilities of success, and so supports the estimation of Rasch measures.

Following Fisher (1922), the likelihood of the data set, L, is the product of the probabilities of the data points:

$$L = \prod_{n,i} P_{nix}. \quad (4)$$

Non-iterative Estimation Methods

Rasch measures are additive, and so linear. Ordinal data are of unknown linearity. This means that Rasch estimates are non-linear transformations of data. Usually, estimation with non-linear functions requires an iterative approach, in which initial rough estimates are systematically improved until final estimates are obtained. There are, however, two estimation methods which do not require iteration.

Graphical methods

When all items are of equal difficulty, D, then B_n can be estimated in closed form:

$$B_n = \log\left(\frac{P_{i1}}{P_{i0}}\right) + D \approx \log\left(\frac{R_n}{L - R_n}\right) + D. \quad (5)$$

Similarly when all persons are of equal ability, B, then D_i can also be estimated in closed form:

$$D_i = B - \log\left(\frac{P_{i1}}{P_{i0}}\right) \approx B - \log\left(\frac{R_i}{N - R_i}\right). \quad (6)$$

Of course, both these conditions cannot hold simultaneously. Nevertheless, for rough approximations when precise measures are not required, these logistic transformations of raw scores provide the basis for a simple graphical method.

Georg Rasch (1960, Ch. VI) demonstrates how plotting logistic transformations of success frequencies permits the drawing of trace lines by eye. The persons are stratified by raw score, R, on the complete test. Each raw score is converted into a logit, log(R/(L-R)). For each score group, their percent success on each item is computed and converted into its logit value, log(success%/failure%). For each item, the success logits are plotted against the score logits.

Figure 1 shows the empirical jagged success-logits for each item and person score-group, together with the inferred parallel straight Rasch

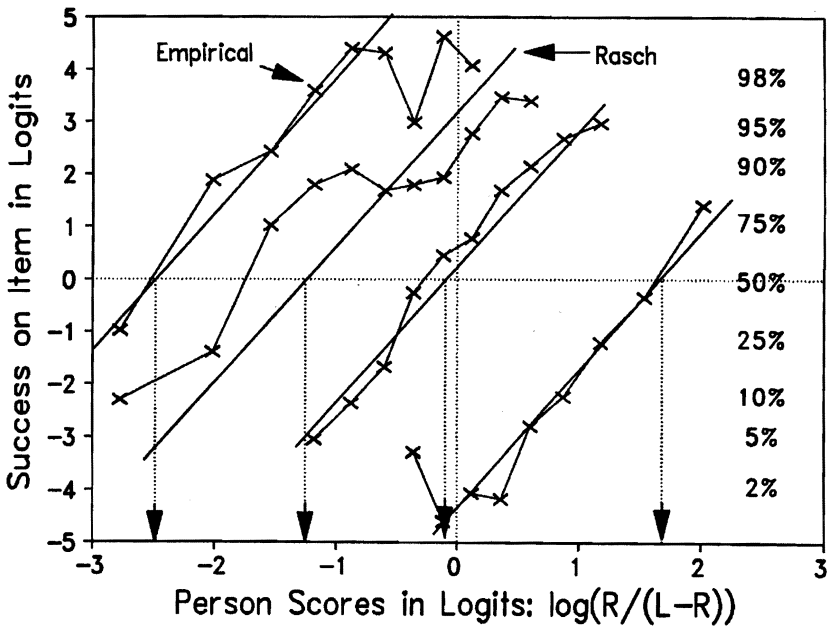


Figure 1. Graphical estimation of the difficulty measures of four items.

lines, drawn by eye. This Figure reports 4 items from G. Rasch's BPP test results. In the plot, the difficulty of an item is equal to the ability of people whose probability of success is 0.5. Though these estimates are somewhat compressed and distorted, they are not misleading as to the hierarchy and relative placement of persons and items on the latent variable. The plot also supports an investigation into measure accuracy in terms of the fit of the data to the Rasch model. Whenever empirical raw-score-based item characteristic curves are available, logistic transformation of both axes yields plots equivalent to those produced by G. Rasch.

Non-iterative Normal Approximation estimation (PROX)

The graphical method failed to allow for differences among person abilities and item difficulties. Leslie Cohen (1979) deduced a non-iterative method, PROX, for estimating Rasch measures, when the data are complete and both items and persons are approximately normally distributed. A procedure for performing PROX by hand is given in Wright and Stone (1979, Chap.2). Even when the distributional assumptions are not met, PROX provides useful starting values for the other estimation methods.

There is a convenient arithmetical relationship between the unit-normal ogive and the logistic ogive. Berkson (1944) takes advantage of it for bio-assay calculations. The relationship between the ogives is specified as:

$$\Psi^{-1}(y) \approx 1.7\Phi^{-1}(y), \tag{7}$$

where Ψ is the logistic function and ϕ is the normal cumulative function The standard equating value of 1.7 minimizes the maximum difference between the functions across their whole range (Camilli, 1994). Linacre (1997a) suggests 1.65 as a better equating value for Rasch use.

When dichotomous data are complete and the parameters of each facet approximate a normal distribution, then non-iterative estimation equations are:

$$B_n = \sum_{i=1}^{L} D_i + X_D \log\left(\frac{R_n}{L - R_n}\right), and \tag{8}$$

$$D_i = C_D - X_B \log\left(\frac{R_i}{N - R_i}\right). \tag{9}$$

Since, by convention, $\Sigma D_i = 0$ establishes the local origin of the measurement scale, C_D is chosen so that $\Sigma D_i = 0$.

X_D and X_B are obtained from S_D and S_B, the population raw-score-based standard deviations. S_D the standard deviation of success on the test items is given by

$$S_D = S.D.\left(\log\left(\frac{R_i}{L - R_i}\right)\right), where\,(i = 1, L). \tag{10}$$

S_B the standard deviation of the success of the person sample is given by

$$S_B = S.D.\left(\log\left(\frac{R_n}{L - R_n}\right)\right), where\,(n = 1, N). \tag{11}$$

X_D adjusts for test width. The wider the spread of success on test items, S_D, the wider the spread of persons being measured, and so the bigger the measure difference between a low scoring and a high scoring person.

$$X_D = \left(\frac{1 + S_D / 2.89}{1 - S_D S_B / 8.35}\right)^{\frac{1}{2}} \tag{12}$$

X_B adjusts for sample spread. The wider the spread of success by the sample, S_B, the wider the range of item difficulties.

$$X_B = \left(\frac{1 + S_B / 2.89}{1 - S_D S_B / 8.35}\right)^{\frac{1}{2}} \tag{13}$$

Iterative Estimation Methods

Iterative estimation methods adopt initial rough starting values, e.g, zeroes, for the estimates. These estimates are used to obtain expected values for the data. A comparison is made between what was observed and what is expected. Then better estimates are produced which minimize discrepancies. This process is repeated, i.e., iterated, until the discrepancies are deemed inconsequential. At this point, the estimation process has converged.

Most estimation methods employ some form of the method of maximum likelihood. The goal of this method, due to Fisher (1922), is to discover the parameter values which maximize the likelihood of the data,

under whatever constraints the analyst imposes. An advantage of the method is that, in general, a second derivative of the likelihood function provides a standard error for the estimate.

The choice of constraints optimizes certain aspects of the estimation process or the estimates themselves, but always at a cost. For instance, there is the ideal of estimation consistency. A consistent estimation procedure produces estimates that asymptotically approach their latent, "true", values as the size of the data set increases. This might appear to be an essential feature of any estimation procedure, but it is not. First, estimation procedures which are consistent according to one method of increasing the data set, can be inconsistent according to another. Second, the inconsistency may be so small as to have no practical implications. Third, the inconsistency in any finite data set, termed "statistical bias", may be correctable. On the other hand, insisting on estimation consistency may prevent estimation under specific conditions, e.g., in the presence of missing data.

In nearly all estimation methods, extreme (zero and perfect) marginal scores imply out-of-range parameter values and so are inestimable.

Table 1

Iterative Estimation Methods

Type	Acronym	Name	Shortcomings	Software
Datum-by-datum		Gaussian Least-Squares	Many measures per score	
		Minimum Chi-Square	Many measures per score	
	PAIR	Pairwise	Many measures per score Correctable standard errors	RUMM
Marginal without distributional assumptions	CMLE CON FCON	Conditional Maximum Likelihood Estimation	Missing data intolerant Limited analysis size	Lpcm-Win
	JMLE UCON	Joint Maximum Likelihood Estimation	Correctable statistical bias	Facets Quest Winsteps
		Log-linear	Missing data intolerant	LOGIMO
Marginal with distributional assumptions	MMLE	Marginal Maximum Likelihood Estimation	Distributional mismatches	ConQuest (IRT)
	PROX	Normal Approximation Estimation	Distributional mismatches	Winsteps
		Items Two-at-a-time	Only for desperate situations	

Accordingly, data corresponding to extreme scores must be eliminated before estimates are produced. There are separate estimation techniques for imputing reasonable measures to extreme scores, once the measures for non-extreme scores have been estimated.

Iterative estimation methods are classified here according to several major considerations. (i) Is estimation conceptualized as proceeding datum by datum, or at the marginal (raw score per parameter) level? (ii) Are all parameters estimated or are some conditioned out of the estimation? (iii) Are parameters free or are they modeled as part of a distribution?

Estimation datum-by-datum

A Rasch model resembles a simple form of a "transition odds" or "adjacent logit" logistic (logit-linear) regression model. George Udny Yule (1925) and Joseph Berkson (1944) suggest methods for estimating the parameters of a logistic curve. Several of their methods are generally applicable. These methods are generally robust against missing data.

Gaussian least-squares. This estimation method minimizes the sum of the squares of the differences between what is observed and what is expected across the data. The function to be minimized, F, is:

$$F = \sum_{Data} (X_{ni} - E_{ni})^2. \qquad (14)$$

This must be minimized for all parameters simultaneously. From the perspective of a particular parameter, say B_n, the minimization occurs when:

$$\frac{dF}{dB_n} = \sum_{n} 2(E_{ni} - X_{ni})V_{ni} = 0, \qquad (15)$$

where V_{ni} the model variance of an observed rating about its expectation, is

$$V_{ni} = (1 - E_{ni})^2 P_{ni1} + (0 - E_{ni})^2 P_{ni0} = P_{ni1} P_{ni0}. \qquad (16)$$

From the Rasch measurement perspective, a drawback to this and all other datum-by-datum methods is that different response strings with the same total raw score produce different measures.

Minimum chi-square. In contrast to the previous method which minimizes numerical distances on the ordinal scale, the minimum chi-square method maximizes the fit of the data to the Rasch model. Consequently, outlying unexpected observations (such as coding errors) are more influ-

ential in the minimum chi-square approach and, again, the same marginal score can produce different estimates. The function, F, to be minimized for all parameters simultaneously is:

$$F = \sum_{Data} \frac{(X_{ni} - E_{ni})^2}{V_{ni}}. \tag{17}$$

Pairwise estimation (PAIR). Since the Rasch model is a log-odds model, an attractive approach is to use the relative frequencies of observations in the data to estimate the parameters. Suppose that two persons, n and m, respond to the same items. C_{10} is the number of times that person n succeeds on items that person m fails, and *vice-versa* for C_{01}. Then, an estimate of the difference in ability between n and m is given by the paired comparison

$$B_n - B_m \approx \log\left(\frac{C_{10}}{C_{01}}\right). \tag{18}$$

Following this approach, one data set yields estimates of the relative abilities of every pair of persons. These pairs of abilities, however, are likely to be somewhat contradictory. The resolution of these contradictions is to combine the paired comparisons into a likelihood function (Wright and Masters, 1982), which is maximized when the parameter estimates simultaneously satisfy the relationship

$$\frac{dF}{dB_n} = \sum_{m \neq n}^{N} \left(C_{10} - \frac{C_{10} + C_{01}}{1 + e^{(B_m - B_n)}} \right). \tag{19}$$

Within one estimation equation the same observation may be used many times. For instance, if person n is the only person to succeed on item 1, then that success is included in every C_{10} term for person n, and so adds N-1 to the total summation. This multiple use of observations means that standard errors derived from second derivatives are too small, roughly in proportion to the square-root of the average number of times each observation is used.

In this method, one set of parameters, usually the item difficulties, are estimated. Then another set is estimated using the pairwise method and the two sets of estimates are aligned on one measurement continuum. Alternatively, the pairwise estimates are set as fixed values (anchors), and another method is used to estimate the other measures from them.

Marginal Estimation without Distributional Assumptions

In marginal models, identical total raw scores, obtained under the same conditions, estimate identical Rasch measures, regardless of the specifics of the response string. This accords with Fisher's (1922) concept of sufficiency, but has been deemed counter-intuitive by empiricists. In general, however, any argument proposing that getting a hard item unexpectedly correct merits a higher measure can be offset by an equivalent argument that getting an easy item unexpectedly wrong merits a lower measure.

Item Response Theory (IRT) models generally require assumptions about the distribution of the latent parameters in order to be estimable. Rasch parameters, however, can be estimated with or without distributional assumptions regarding the parameters. There is one distributional specification, however, that is deemed to hold across these estimation methods. The unmodeled part of each datum, the residual difference between the observed and the expected values, is specified to be normally distributed, when the residual is standardized by its own model variance.

Conditional Maximum Likelihood Estimation (CMLE). This methods capitalizes on the proposition that identical person raw scores produced under identical conditions imply identical measures, but avoids actually estimating those measures. This is achieved by stratifying the person sample by raw score, and them estimating item difficulties within each raw score stratum. Estimation within stratum conditions out the person measures, resulting in estimates with minimal statistical bias and well-defined standard errors.

The minimal remaining estimation bias results from the very slight probability that a sample of respondents, whose measures correspond to the estimated parameters, would all simultaneously succeed or fail on a test item. If a large sample of respondents is obtained, then this probability is effectively zero, meaning that CMLE estimates become free of bias.

Estimation is conceptually simple, but challenging to implement. First, a set of rough starting values for the item difficulties is imputed. Then the likelihood of every possible response string that generates a particular score, r, is estimated. In this computation, a reference person of any computationally convenient ability can be used. All these likelihoods are summed for the particular score, r, becoming the likelihood of making that score in any way, Λ_r. The response strings for score r are then inspected for each item in turn. The likelihoods of all response strings with a success on item i are accumulated into a likelihood of observing a success on item i given a score of r, Λ_{ri}. In score stratum r, there are N_r persons. Thus the expected number of suc-

cess on item i in score stratum r is $N_r(\Lambda_{ri}/\Lambda_r)$. Finally, a revised estimate of the difficulty of item i is obtained by summing across all score strata and applying an estimation equation like

$$D_i' = D_i - \frac{R_i - \sum_{r=1}^{L-1} N_r \frac{\Lambda_{ri}}{\Lambda_r}}{(ModelVariance)}. \tag{20}$$

There is only a limited number of exponential terms that can be summed into Λ_r without loss of computational precision. This has restricted CMLE to short tests. Improvements in computer hardware and more sophisticated numerical methods have aided CMLE (Verhelst and Glas, 1995), but it is still impractical for long tests or test with many different patterns of missing data.

Joint Maximum Likelihood Estimation (JMLE). The Rasch measures for which the data are most likely to be observed are those for which the observed and expected scores coincide. Since raw scores are sufficient statistics for both items and persons measures, all measures can be estimated simultaneously. In JMLE, no parameters are conditioned out, so the method is also termed "unconditional" (UCON).

Since the marginal scores coincide with their expectations, JMLE estimates satisfy the optimal least squares criterion,

$$\left(R_n - \sum_{i=1|n}^{L} E_{ni} \right)^2 = 0. \tag{21}$$

As in all these methods, the final estimates are independent of the iterative path followed, but the usual approach follows Newton-Raphson. The estimation equation to produce a better estimate B_n' of the previous estimate B_n is:

$$B_n' = B_n + \frac{R_n - \sum_i E_{ni}}{\sum_i V_{ni}}, \tag{22}$$

where V_{ni} is defined in (16). This estimation method has proved robust against missing data, and also easily allows the incorporation into one analysis of data generated by variants of the Rasch model (dichotomous, partial credit, rating scale, Poisson, etc.).

A long-standing criticism of this method is that it is prone to noticeable estimation bias with short tests. For instance, if a two item dichotomous test were given to a sample of persons, the estimated difference between the item measures according to JMLE would be twice that estimated by the pairwise estimation method. In practice, however, this bias has few implications because the relative ordering and placement of the estimates is maintained. When JMLE is used to estimate measures from paired comparison data, a correction factor of 0.5 removes the statistical estimation bias (Linacre,1997b).

JMLE is amenable to pre-set (fixed, anchored) parameter estimates, so that it is often used to estimate those parameters which have been left unestimated by other estimation methods.

Log-linear methods. Logit-linear (logistic) models, including Rasch models, can be reparameterized as log-linear models and applied to frequency tables. In the frequency table, there is a cell for each observed, non-extreme, pattern of responses to the items. The cell contains a count of the number of times the pattern is observed. Then, for a particular response string s with frequency F_s,

$$\log(F_s) = -\sum_s X_{si} D_i + \gamma, \qquad (23)$$

where $X_{si} = 1$ if the response to item i in string s is 1, and $X_{si} = 0$ otherwise. The item difficulties, D_i, are estimated, and γ is chosen such that $\Sigma D_i = 0$. Person abilities can be estimated by another method, after anchoring the item difficulties.

The parameters of log-linear models can be estimated with standard statistical computer programs, but these have proved of limited utility. Rasch models, in log-linear form, can have hundreds of parameters and millions of cells. These overwhelm the computational capacity of most statistical software. Further, sometimes the design of the estimation algorithm requires a cell for every possible response string. Then many of the cells will contain incidental zeroes, because particular response strings did not happen to be observed. Further, if data are missing from response strings, estimation with standard statistical software becomes virtually impossible.

Kelderman (1984) has devised an estimation approach specifically for Rasch log-linear models. This is implemented in Kelderman and Steen (1988). It can handle long tests, but is still intolerant of missing data.

Marginal Estimation with Distributional Assumptions

Distributional assumptions regarding some or all of the parameters can be usefully employed to simplify computation or even make estimation possible. If the distributional assumptions seriously mismatch the latent parameter distributions, then severe estimation bias may be introduced.

Marginal Maximum Likelihood Estimation (MMLE). MMLE imposes a distribution function on the subject parameters. The simplest function is a normal distribution (paralleling IRT estimation). More sophisticated functions are also employed such as multivariate normal distributions based on demographic variables (Adams, et al., 1977) and empirical-Bayesian distributions.

MMLE can surmount several obstacles at which other estimation methods balk. First, it permits the estimation of sample measure characteristics even when there is insufficient information to produce meaningful estimates for individuals within the samples. In particular, extreme scores, very short response strings, Guttman patterns and missing data can be easily managed. Second, when the intention is not to measure individuals, but to summarize estimates, it bypasses an analytic step. Third, it supports generalized multidimensional forms of the Rasch model (Wu, et al., 1998).

MMLE produces estimates for the discrete parameters, usually corresponding to item difficulties, such that their observed and expected marginal scores coincide, under the condition that the distribution of the other parameters has the required form. This requires a two-stage estimation approach, such as the E-M, Expectation-Maximization, algorithm (Bock and Aitken, 1981).

The MMLE method is ubiquitous in the estimation of the two- and three-parameter Item Response Theory (IRT)models. When those models are constrained to take the form of Rasch models, then Rasch MMLE estimates are obtained.

Normal Approximation Algorithm (PROX). The PROX algorithm has an iterative form which can accommodate missing data. The estimation equations are resubscripted to indicate that only those instances when person n actually responded to item i are to be considered. For instance,

$$B_n = \sum_{i \in n}^{L_n} D_i + X_{D_n} \log\left(\frac{R_n}{L_n - R_n}\right), \tag{24}$$

where ΣD_i applies only to those L_n items encountered by person n. X_{Dn} refers to the spread of those items. Linacre (1994) derives iterative PROX estimation equations for missing data. Linacre (1995) extends PROX to polytomous data.

Items two-at-a-time

When tests are short, many subjects obtain extreme scores. These introduce an unquantifiable amount of bias into summary statistics. The focus of measurement, however, may not be the subjects, but the samples to which they belong. When subjects are regarded as normally distributed, Wright (1998b) suggests estimation equations for the sample mean and standard deviation from the responses of subjects to pairs of items.

Imagine that a large sample of people have taken two dichotomous items, A and B, approximately as the Rasch model predicts. Table 2 is the tabulation of their scored responses. According to the Rasch model, the difference between the item difficulties is estimated directly by

$$D_A - D_B \approx \log\left(\frac{N_{01}}{N_{10}}\right), \quad with\ S.E. \approx \sqrt{\frac{N_{10}+N_{01}}{N_{01}N_{10}}}. \tag{25}$$

If we assume that the sample is normally distributed, then we can estimate the sample mean and standard deviation. The sample mean ability is relative to the average difficulty of the two items. A simulation study reported in Wright (1998b) suggests the following estimator:

$$Sample\ Mean \approx 1.864\left[\log\left(\frac{T_{A1}}{T_{A0}}\right)+\log\left(\frac{T_{B1}}{T_{B0}}\right)\right]+1.455\log\left(\frac{N_{00}}{N_{11}}\right). \tag{26}$$

Table 2

Counts on a Two-Item Test

		Item B		
		Right: 1	Wrong: 0	Totals:
Item A	Right :1	N_{11}	N_{10}	T_{a1}
	Wrong: 0	N_{01}	N_{00}	T_{A0}
Totals:		T_{B1}	T_{B0}	T

An estimator for sample standard deviation is:

$$S.D. \approx 3.763 + 1.4 * \left[\log\left(\frac{N_{11}}{T - N_{11}}\right) + \log\left(\frac{N_{00}}{T - N_{00}}\right) \right] + 0.0101 * \log\left(\frac{N_{10}}{N_{01}}\right)^2$$
$$+0.081 \left[\log\left(\frac{T_{A1}}{T_{A0}}\right)^2 + \log\left(\frac{T_{B1}}{T_{B0}}\right)^2 \right] \quad (27)$$

Estimating by Other Methods and for Other Models

Estimation methods for dichotomous data are further discussed in Molenaar (1995), and Hoijtink and Boomsma (1995), generally in a context of short tests with no missing data. Andrich (1988, Chap. 5) provides worked examples of CMLE, JMLE and PAIR in a broader context.

Most of estimation methods have been broadened to cover polytomous and other extended Rasch models. The characteristics of the estimation methods remain the same. Wright and Masters (1982) provide algorithms for CMLE, JMLE and PAIR. Andersen (1995) addresses CMLE and MMLE.

Estimating Extreme Scores

Under strict Rasch model conditions, extreme (zero and perfect) scores correspond to indefinite measures, and can take any value outside the measurement range of the test. Consequently, under most estimation methods, the response vectors corresponding to extreme scores are dropped from the analysis. In many situations, however, measures must be reported for extreme scores, or the measures corresponding to extreme scores must be included in summary statistics.

There are two approaches to imputing measures for extreme scores. The first approach is to consider extreme scores to be part of a measure distribution. This requires an estimation method, such as MMLE, that estimates at the sample, rather than individual, level. The second approach is to apply some reasonable inference about the nature of the extreme score, and use this to estimate a measure.

Wright (1998a) suggests nine bases for choosing a measure corresponding to an extreme score. He concludes that, for dichotomous data, reasonable measures for extreme scores are between 1.0 and 1.2 logits more extreme than the measures for the most outlying non-extreme scores.

For polytomous data, measures corresponding to scores between 0.25 and 0.5 score-points more central than the extreme scores can be usefully imputed as the extreme measures.

Estimation Error

A recurring theme in the literature of the Rasch model is estimation error. No estimation technique can guarantee to reproduce the exact measures of the generating parameters, even when the data fit the Rasch model. The difference between the estimates and the generators is termed estimation error. There are three main sources of estimation error: deficiencies in the theoretical properties of the estimates, deficiencies in the implementation of the estimation algorithm and mismatches between the distribution of the data and the assumptions of the estimation algorithm.

Some techniques could recover the generators, in theory, if they were provided infinite data of the right kind. For instance, the "two-at-a-time" and pairwise estimation techniques would recover the exact measure difference between items, given the responses of an infinite number of on-target persons under Rasch model conditions. Such estimation techniques are termed "consistent". Though a desirable property, consistency is not of practical concern.

A theoretical deficiency in most estimation methods causes some degree of estimation bias, which can noticeably affect measures estimated from short tests or with small samples. Even then, the bias can usually be easily corrected (Wright, 1988). An example is the correction of bias in measures resulting from the use of JMLE for analyzing measures from paired-comparison observations (Linacre, 1984). Under Rasch model conditions, estimation bias is due to the inclusion of the possibility of extreme score vectors in the computations of the estimation algorithms, even though they must be eliminated from the data (or other arbitrary constraints introduced), because they produce infinite parameter estimates.

The bias in JMLE is chiefly caused by the likelihood of persons obtaining extreme scores. Linacre (1989) derives a JMLE-based estimation algorithm (XCON) which overcomes this deficiency, but there has been no demand, as yet, to implement it in a generally accessible way. CMLE is relatively bias free, because person extreme scores are eliminated from the estimation space, and there is only a remote possibility of an extreme score for an item.

Deficiencies in implementing estimation algorithms are most apparent with CMLE. Computations of the likelihoods of every possible re-

sponse string that generates each observed raw score is required. This is a large computational load and, worse, involves the accumulation of many small numbers. Loss of numerical precision can result, leading to error in the estimates.

Mismatches between the distributional assumptions of the estimation algorithm and the data can skew MMLE and PROX estimates. PROX capitalizes on the normal distribution, so that good estimates will not be obtained with a highly skewed sample, such as those found in many clinical situations. MMLE can use more sophisticated methods to model the observed parameter distribution, but the match is always approximate.

Standard Errors of Measures

It is impossible to obtain point-estimates of Rasch parameters. Like all other measures, every Rasch measure is to some degree imprecise. This imprecision is usually reported as a standard error. For MMLE, it may be reported as a series of plausible values, intended to report a more complex error distribution, but, for practical purposes, even these can be summarized by a mean (corresponding to the estimate) and a standard deviation (corresponding to the standard error).

The algorithm to compute the standard error is derived from the properties of the estimates or is a by-product of the estimation method. The pairwise standard error is less well-defined than those of the other estimation methods because of the data-dependent reuse of observations in estimating observations. Correcting for the degree of data reuse results in serviceable standard errors.

All estimation methods produce estimates with standard errors of about the same size, because they are obtained from data containing the same information. In general, the more observations in which a parameter participates, the smaller the standard error of its estimate. The information in an individual observation is most influenced by the targeting of the parameters that generated the observation and the number of categories in the relevant rating scale. Covariance in the data reduces precision. Adjustment for covariance inflates the standard errors, but rarely to the extent that it would lead the analyst to a substantively different conclusion about the quality of the measures.

Regardless of the estimation method, there are four conventional ways of reporting Rasch standard errors (Wright, 1995). Standard errors can either be local or general. They can also be ideal or real.

Most Rasch estimation programs report local, ideal standard errors. JMLE estimates are usually characterized with general standard errors.

Local standard errors are computed relative to the estimate of some particular item on the test (usually the first one). This reference item has no standard error. Choice of a different reference item changes all the standard errors. This makes the standard errors difficult to interpret and awkward to transport to other contexts.

General standard errors are computed as though all other parameters are known, i.e., as though their estimates are point-estimates. Converting from general to local standard errors is merely a matter of choosing a reference item, and then computing joint standard errors between that reference item and all other items. The general standard errors have the virtue that they are easy to interpret and transport to other contexts.

Ideal standard errors reflect the highest possible precision obtainable with data like those observed. These "best case" values are the smallest possible, estimated on the basis that the data fit the Rasch model. Any idiosyncracies in the data are regarded merely as evidence of the stochastic nature of the model. These "model" standard errors produce the highest possible estimates of test reliability.

Real standard errors reflect the most imprecision. These "worst case" values are obtained on the basis that all idiosyncracies in the data are contradictions to the Rasch model. These values will produce the lowest reasonable estimates of test reliability. As misfit in the data is brought under control, the real standard error approaches the ideal.

Implementations of the Estimation Methods

Rasch estimation methods are rarely implemented directly by the data analyst, except perhaps for the estimation of person measures when item difficulties are known (Linacre, 1996, 1998). Instead, analysts rely on available computer programs.

To illustrate the similarities between the estimates obtained by different estimation approaches, five computer programs were used. RUMM (Andrich, et al., 1997) implements pairwise estimation. Quest (Adams and Toon, 1994) and Winsteps (Wright and Linacre, 1991) implement JMLE. ConQuest (Wu, et al., 1998) implements MMLE. Lpcm-Win (Fischer, 1998) implements CMLE.

Though the intention was to analyze the same data set, representative of actual clinical data, with all 5 programs, this proved impossible

with the versions of the programs available to the author. Instead, two data sets were used. One data set comprised 16 items and 156 persons. The items were polytomous with up to 4 categories. The data set included extreme scores and missing data. It was provided as a sample data set with the RUMM program. Measures were estimated from this data set with ConQuest, Quest, RUMM and Winsteps. A second data set was constructed from this data set. It comprised 15 items and 156 persons. There were no extreme scores nor missing data. Measures were estimated from this data set with Lpcm-Win, ConQuest and Winsteps.

Each computer program was instructed to produce estimates in accordance with the Rasch partial credit model, but using the program's own default settings, as far as possible. Every estimation process was continued to convergence. Item, rating scale and person estimates were produced, to the extent each program allowed.

On inspection of program output, it was seen that item difficulties and rating scale (partial credit) estimates were reported in such different ways that simple comparison was not possible. It also emerged that there were two ways of reporting person measures, either case-by-case or for all possible non-extreme scores with no missing data. The information provided by these two ways is combined for this discussion. Since most programs did not attempt to estimate measures corresponding to extreme scores, these are not considered here.

Figure 2 depicts the person measures produced by four of the programs on the first data set. Though the programs themselves adopt different criteria for establishing the local origin of the measurement scale, all measures are equated to a common local origin in the Figure. Winsteps was run in its default mode which does not attempt to correct for JMLE estimation bias. This bias causes its estimates (represented by the diagonal) to be slightly wider (less central) than those of the other programs. It appears that Quest, also using JMLE, is correcting for estimation bias. The standard errors of the measures in this plot are 0.4 logits. All four programs, and so all four estimation methods, are producing substantively and statistically the same measures.

Figure 3 plots person measures estimated from the second data set. At the lower end, the estimates coincide. For these estimates, standard errors are again 0.4 logits. At the upper end, differences are seen. Winsteps produced JMLE estimates without correction for estimation bias, represented by the diagonal line, the highest estimates. The Lpcm-Win (CMLE) measures are next most central, plotted as X. The ConQuest (MMLE) measures

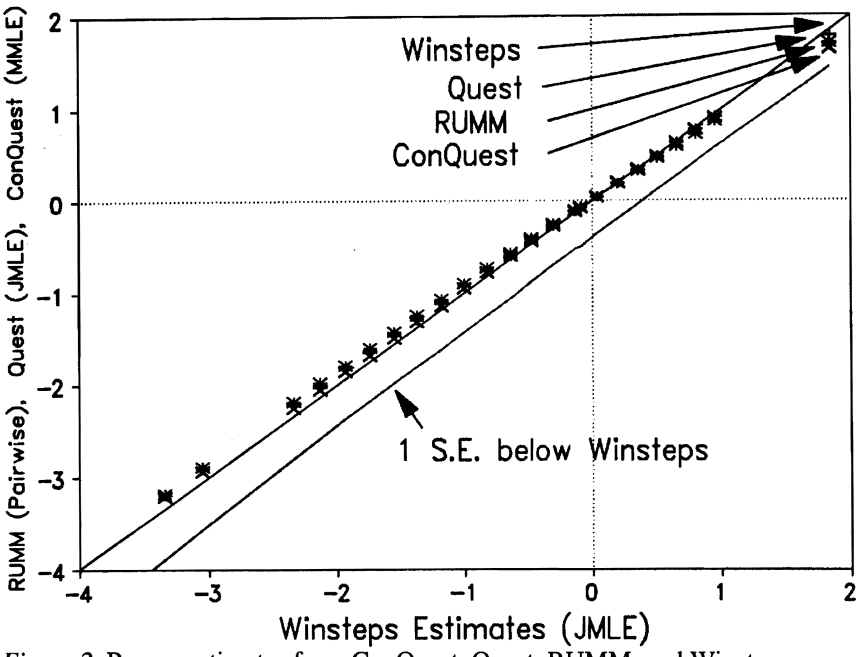


Figure 2. Person estimates from ConQuest, Quest, RUMM, and Winsteps.

are the most central, plotted as +. The range of estimates of the most extreme person in the top right of the Figure is 0.7 logits. A line one standard error below the Winsteps estimates is also plotted. Again it is seen that the estimates are statistically identical. Confusion might result, however, if measures from one program were interspersed with those from another.

Conclusion

Each Rasch estimation method has its strong points and its advocates in the professional community. Each also has its shortcomings. Nevertheless, when the precision and accuracy of estimates are taken into account (Wright, 1988), all methods produce statistically equivalent estimates. Care needs to be taken, however, when estimates produced by different computer programs or estimation methods are to be compared or placed on a common measurement continuum.

Acknowledgment

Andrew Stephanou of the Australian Council for Educational Research provided valuable suggestions.

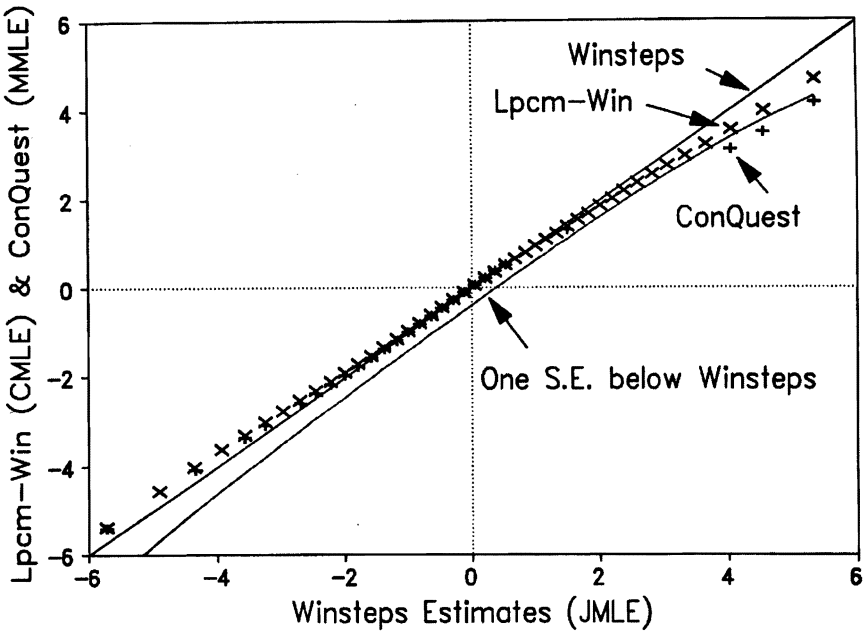


Figure 3. Person estimates from ConQuest, Lpcm-Win and Winsteps.

References

Adams, R. J., and Toon, K. S. (1994). *Quest: The Interactive Test Analysis System.* Melbourne, Australia: Australian Council for Educational Research.

Adams, R. J., Wilson, M. R., and Wu, M. L. (1997). Multilevel item response models: an approach to errors in variable regression. *Journal of Educational and Behavioral Statistics, 22* (1), 46-75.

Andersen, E. B. (1995) Polytomous Rasch models and their estimation. Chapter 15 in G. H. Fischer and I. W. Molenaar, *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer Verlag.

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

Andrich, D. A., Lyne, A., Sheridan, B., Luo, G. (1997). *RUMM: Rasch Unidimensional Measurement Models.* Perth, Australia: RUMM Laboratory.

Berkson, J. (1944). Applications of the logistic function to bio-assay. *Journal of the American Statistical Society 39,* 357-365

Bernoulli, J. (1713). *Ars Conjectandi. Part 4.* Basel. Excerpted in *Rasch Measurement Transactions, 12* (1), 625. 1998.

Bock, R. D. and Aitken, M. (1981) Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika, 46,* 443-459.

Camilli, G. (1994). Origin of the scaling constant d=1.7 in item response theory. *Journal of Educational and Behavioral Statistics 19* (3), 293-5.

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology 32* (1), 13-120.

Fischer, G. H. (1995). Derivations of the Rasch model. Chapter 2 in G. H. Fischer and I. W. Molenaar, *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer Verlag.

Fischer, G. H. (1998). *Lpcm-Win.* Minneapolis, MN: Assessment Systems Corp.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Proceedings of the Royal Society, 222*, 309-368.

Hoijtink, H., and Boomsma, A.(1995). On person parameter estimation in the dichotomous Rasch model. Chapter 4 in G. H. Fischer and I. W. Molenaar, *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer Verlag.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49*, 223-245.

Kelderman, H., and Steen, R. (1988). *LOGIMO computer program for log-linear item response theory modelling.* Twente, The Netherlands: University of Twente.

Laudan, L. (1977). *Progress and its Problems*. Berkeley, CA: University of California Press.

Linacre, J. M. (1984). Paired comparisons with standard Rasch software. *Rasch Measurement Transactions, 13*(1), 584-5.

Linacre, J. M. (1989). *Many-facet Rasch Measurement.* Chicago: MESA Press.

Linacre, J. M. (1994). PROX with missing data. *Rasch Measurement Transactions, 8*(3), 378.

Linacre, J. M. (1995). PROX for polytomous data. *Rasch Measurement Transactions, 8*(4), 400-401.

Linacre, J. M. (1996). Estimating measures with known item difficulties. *Rasch Measurement Transactions, 10*(2), 499.

Linacre, J. M. (1997a). The normal cumulative distribution and the logistic ogive. *Rasch Measurement Transactions, 11*(2), 569.

Linacre , J. M. (1997b) Paired comparisons with standard Rasch software. *Rasch Measurement Transactions, 11*(3), 584-5.

Linacre, J. M. (1998). Estimating measures with known polytomous item difficulties. *Rasch Measurement Transactions, 12*(2), 638.

Molenaar, I. W. (1995). Estimation of item parameters. Chapter 3 in G. H. Fischer

and I. W. Molenaar, *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer Verlag.

Rasch, G. (1960) Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: University of Chicago Press. Reprinted, 1992. Chicago: MESA Press.

Verhelst, N. D., and Glas, C. A. W. (1995). The one parameter logistic model. Chapter 12 in G. H. Fischer and I. W. Molenaar, *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer Verlag.

Wright, B. D. (1988). The efficacy of unconditional maximum likelihood bias correction: comment on Jansen, Van den Wollenberg, and Wierda. *Applied Psychological Measurement, 12*, 315-318.

Wright, B. D. (1995). Which standard error? *Rasch Measurement Transactions, 92)*, 436-7.

Wright, B. D., (1998a). Estimating measures for extreme scores. *Rasch Measurement Transactions, 12*(2), 632-633.

Wright, B. D. (1998b). Two-item testing. *Rasch Measurement Transactions, 12*(2), 627-8.

Wright, B. D., and Linacre, J. M. (1991). *Winsteps Rasch Measurement Computer Program.* Chicago: MESA Press.

Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

Wright, B. D., and Stone, M. H. (1979). *Best Test Design.* Chicago: MESA Press.

Wu, M. L., Adams, R. J., Wilson, M. R. (1998). *ConQuest: Generalised Item Response Modelling Software.* Melbourne, Australia: Australian Council for Educational Research.

Yule, G. U. (1925). The growth of population and the factors which control it. Presidential address. *Journal of the Royal Statistical Society, 88*, 1-62.

Journal of Outcome Measurement®

Volume 3

Author and Title Index

CONTRIBUTOR INFORMATION

Content: *Journal of Outcome Measurement* publishes refereed scholarly work from all academic disciplines relative to outcome measurement. Outcome measurement being defined as the measurement of the result of any intervention designed to alter the physical or mental state of an individual. The *Journal of Outcome Measurement* will consider both theoretical and applied articles that relate to measurement models, scale development, applications, and demonstrations. Given the multi-disciplinary nature of the journal, two broad-based editorial boards have been developed to consider articles falling into the general fields of Health Sciences and Social Sciences.

Book and Software Reviews: The *Journal of Outcome Measurement* publishes only solicited reviews of current books and software. These reviews permit objective assessment of current books and software. Suggestions for reviews are accepted. Original authors will be given the opportunity to respond to all reviews.

Peer Review of Manuscripts: Manuscripts are anonymously peer-reviewed by two experts appropriate for the topic and content. The editor is responsible for guaranteeing anonymity of the author(s) and reviewers during the review process. The review normally takes three (3) months.

Manuscript Preparation: Manuscripts should be prepared according to the *Publication Manual of the American Psychological Association* (4th ed., 1994). Limit manuscripts to 25 pages of text, exclusive of tables and figures. Manuscripts must be double spaced including the title page, abstract, text, quotes, acknowledgments, references, and appendices. On the cover page list author name(s), affiliation(s), address(es), telephone number(s), and electronic mail address(es). On the second page include a 100 to 150 word abstract. Place tables on separate pages. Include photocopies of all figures. Number all pages consecutively.

Authors are responsible for all statements made in their work and for obtaining permission from copyright owners to reprint or adapt a table or figure or to reprint a quotation of 500 words or more. Copies of all permissions and credit lines must be submitted.

Manuscript Submission: Submit four (4) manuscript copies to Richard M. Smith, Editor, *Journal of Outcome Measurement*, Rehabilitation Foundation Inc., P.O. Box 675, Wheaton, IL 60189 (e-mail: jomea@rfi.org). Prepare three copies of the manuscript for peer review by removing references to author(s) and institution(s). In a cover letter, authors should indicate that the manuscript includes only original material that has not been previously published and is not under review elsewhere. After manuscripts are accepted authors are asked to submit a final copy of the manuscript, original graphic files and camera-ready figures, a copy of the final manuscript in WordPerfect format on a 3 ½ in. disk for IBM-compatible personal computers, and sign and return a copyright-transfer agreement.

Production Notes: Manuscripts are copy-edited and composed into page proofs. Authors review proofs before publication.

SUBSCRIBER INFORMATION

Journal of Outcome Measurement is published four times a year and is available on a calendar basis. Individual volume rates are $35.00 per year. Institutional subscriptions are available for $100 per year. There is an additional $24.00 charge for postage outside of the United States and Canada. Funds are payable in U.S. currency. Send subscription orders, information requests, and address changes to the Subscription Services, Rehabilitation Foundation, Inc. P.O. Box 675, Wheaton, IL 60189. Claims for missing issues cannot be honored beyond 6 months after mailing date. Duplicate copies cannot be sent to replace issues not delivered due to failure to notify publisher of change of address. Back issues are available at a cost of $12.00 per issue postpaid. Please address inquiries to the address listed above.

Copyright© 1999, Rehabilitation Foundation, Inc. No part of this publication may be used, in any form or by any means, without permission of the publisher. Printed in the United States of America. ISSN 1090-655X.