JOURNAL OF OUTCOME MEASUREMENT®

Volume 2, Number 3 1998

Second International Outcome Measurement Conference Highlights 171

Articles

Factor Structure and Dimensionality of the Multidimensional Health Locus of Control Scales in Measuring Adults with Epilepsy 173
Sarah Gehlert and Chih-Hung Chang

Analyzing Nonadditive Conjoint Structures: Compounding Events by Rasch Model Probabilities 191
George Karabatsos

A Research Program for Accountable and Patient-Centered Health Outcome Measures 222
William P. Fisher

Measuring Individual Differences in Change with Multidimensional Rasch Models 240
Wen-chung Wang, Mark Wilson, and Raymond J. Adams

Detecting Multidimensionality: Which Residual Data-type Works Best? 266
John Michael Linacre

Indexing/Abstracting Services: JOM is currently indexed in the *Current Index to Journals in Education* (ERIC).

Second International Outcome Measurement Conference

Conference Highlights

May 15-16, 1998

International House, University of Chicago

Over 75 participants, including 20 from outside the United States, attended IOMC2. This conference, both in its participants and its approach to health outcomes measurements, was a global event. The 22 invited presenters for this conference pursued a bold agenda systematically and efficiently representing historical, theoretical, practical, and future perspectives on the contemporary movement in the healthcare industry to objective outcomes measurement. Underlying the entire conference was a remarkably consistent emphasis on measurement fundamentals and their importance to the practice of outcomes measurement. Issues of precision, order, and reproducibility permeated the presentations both in applications and theoretical discussions providing a common structure of concepts and language for this quickly maturing field. Some of the most exciting aspects of this conference were the previews on a outcome measurement future that is already here. Implementation of standard measurement practices, calibration of international instruments, portable and interchangeable scales, not to mention super-structured data bases linking measures for incoming and outgoing patients are quickly becoming part of the healthcare scene. Technical advancements such as pivot anchoring were presented for the first time.

This balance between old and new was successfully maintained throughout the conference. Ben Wright, for example, simultaneously emphasized an old idea in measurement, individual person measurement rather than groups, yet urged sensibility on a perennial, but current issue concerning the effective level of precision for an application ("How much precision do you need?). His statement concerning the misuse of correlation coefficients because they are based on the computation of variances not on scale distances, while an old statement, is very new to most practitioners.

The historical perspective was best presented by Carl Granger when he described the ALPHA FIM™ for measuring functional independence, a landmark for outcomes measurement representing the most sophisticated and comprehensive measuring system ever achieved with nonphysical scales. Likewise, Larry Ludlow's presentation of the PEDI is a historical achievement in objectively measuring developmental disabilities. In contrast, John Ware presented a perspective on the next generation of outcome measures that will be imple-

mented through computer adaptive testing. As part of an international health care policy commitment, he revealed a "unified measurement strategy" in which known and common metrics will have an even greater importance in advancing the standardization of health care measures across populations and clinical practices.

The technical debates, as well, were both old and new. Richard Smith and Larry Ludlow summarized the interpretation of Infit and Outfit values when assessing fit because of their regular distributional properties unlike the mean square which is very sample size dependent. Among the most interesting new concepts introduced at IOMC2 was pivot anchoring presented by Rita Bode and Allen Heinemann. Essentially a method to refine the definition of a scale when an instrument consists of items with multiple rating formats, their neurologic example provided convincing evidence of its advantages. This emphasis on scale quality was also Craig Velozo's important message in his presentation on approaches to reducing the length of a vision function scale while preserving its measuring properties. Alan Tennant's description of the dangerous inaccuracies that accompany total score assessments was reinforced by his "scale warning" that scales with redundant items "make the use of total scores and percentages irrelevant to outcome judgments". In other words, using uncalibrated scales with redundant items essentially increases the chances that decision making will be based on observer error rather than change in patient status. Karon Cook provided more evidence of this problem, showing that analog scales are really ordinal in nature. This emphasis on response distributions was revisited in Michael Linacre's presentation where he explicitly described an analysis of the ordinal category structure of the FIM using his new analysis program.

The issue of scale compatibility was another reoccurring theme at this conference and effectively presented by Richard Smith in his description of PECS/FIM equating. This research together with work like John Ware's presentation of the SF-36 Mental Health Scale are establishing the groundwork for a major consolidation of health outcome measures.

Nikolaus Bezruczko
Chicago, Illinois

JOURNAL OF OUTCOME MEASUREMENT®, 2 (3), 173 - 190

Factor Structure and Dimensionality of the Multidimensional Health Locus of Control Scales in Measuring Adults with Epilepsy

Sarah Gehlert
The University of Chicago

Chih-Hung Chang
Rush-Presbyterian-St. Luke's Medical Center

External locus of control has been implicated in the development of psychosocial problems in epilepsy, and adults with epilepsy exhibit scores that are more external than those of the normative sample of the Multidimensional Health Locus of Control (MHLC) scales. Although the MHLC scales has the potential to be quite useful in the assessment and treatment of adults with epilepsy, it has not been assessed psychometrically using data from persons with epilepsy. The present study examined the internal consistency, factor structure, and construct validity of the scales using data from a survey of 143 adults with epilepsy. Results from reliability analysis, confirmatory factor analysis, and Rasch analysis supported the hypothesized three-factor structure of the measure, which was internally reliable and factorially valid.

Requests for reprints should be sent to Sarah Gehlert, The University of Chicago, 969 East Sixtieth Street, Chicago, IL 60637.

The Multidimensional Health Locus of Control (MHLC) scales (Wallston, Wallston, & DeVellis, 1978) was developed to assess locus of control attributions specific to health care, and is the most widely used measure of health locus of control. The MHLC is a revision of the Health Locus of Control scale (HLC; Wallston, Wallston, Kaplan, & Maides, 1976), a unidimensional scale constructed in response to Rotter's (1966) suggestion that situation-specific locus of control scales would be of practical as well as theoretical interest (Wall, Hinrichsen, & Pollack, 1989). The authors' move to a multidimensional scale was inspired by Levenson's (1973) questioning of locus of control as a unidimensional construct and her subsequent development of a non-health-specific multidimensional measure of locus of control (Levenson, 1973, 1974, 1975).

The three subscales of the MHLC measure the degree to which health-related outcomes are perceived to be the result of one's own actions, those of significant others in the environment, or luck or chance (DeVellis, DeVellis, Wallston, & Wallston, 1980). The Internality (IHLC) subscale assesses the degree to which an individual believes that his behavior is responsible for his health or illness; the Powerful Others Externality (PHLC) subscale assesses an individual's beliefs that his health or illness is determined by important figures such as physicians, other health professionals, parents, or friends; and the Chance Externality (CHLC) subscale assesses an individual's belief that his level of health or illness is a function of luck, chance, fate, or uncontrollable factors (Rock, Meyerowitz, Maisto, & Wallston, 1987). Two equivalent forms (A and B) of the instrument are available, each of which consists of three 6-item subscales. A third version containing 12-item subscales was derived by combining Forms A and B. Form C, an 18-item scale first described by Wallston in the late 1980s (Wallston, 1989), was published in 1994 (Wallston, Stein, & Smith, 1994). Unlike Forms A and B, which were constructed deliberately to address generic health behaviors or conditions, Form C is a condition-specific locus of control scale.

The MHLC scales have been used as independent, dependent, and correlational variables in numerous studies investigating various health-related behaviors and groups, including hemodialysis patients (Hatz, 1978); hypertensive college students (Sherwin, 1979); persons with epilepsy (DeVellis, DeVellis, Wallston, & Wallston, 1980); participants in a voluntary smoking cessation program (Shipley, 1981); psychiatric patients (Kucera-Bozarth, Beck, & Lyss, 1982); medical and dental students

(Winefield, 1982); inpatient alcoholic population (Russell & Ludenia, 1983); college students (O'Looney & Barrett, 1983); alcoholics (Russell & Ludenia, 1983); mental disorders (Burish, Carey, Wallston, Stein, Jamison, & Lyles, 1984; Horlick, Cameron, Firor, Bhalerao, & Baltzan, 1984; Ingle, Burish, & Wallston, 1984; Nagy & Wolfe, 1983); cigarette smokers (Coelho, 1985); and rehabilitation patients (Umlauf & Frank, 1986).

A good deal of psychometric data have accumulated on the validity and internal consistency of the MHLC, with somewhat mixed results. Nagelberg (1979) reported alpha reliability coefficients for the three MHLC subscales (6-item version) that ranged from .67 to .77; when Forms A and B were combined to yield 12-item subscales, the alpha reliabilities increased to from .83 to .86. Galanos, Strauss, and Pieper (1994) report internal consistency scores of .73, 58, and .68 for the IHLC, CHLC, and PHLC, respectively, for a sample of elderly persons living in the community. Robinson-Whelen and Storandt (1992) report coefficients of .53, .65, and .62 for a sample of participants of similar age. A study of 152 first-year medical and dental students (Winefield, 1982) found alpha coefficients as low as .49 for the CHLC subscale and .58 for the PHLC subscale. McCallum, Keith, and Wiebe (1988) reported alpha coefficients of from .59 to .76. Authors such as Winefield (1982), O'Looney and Barrett (1983), Coelho (1985), and Umlauf and Frank (1986) have noted problems related to the construct validity of the multidimensional approach.

The predictive validity of the MHLC was assessed by its authors by comparing scores on its subscales to general health status (Wallston, Wallston, & DeVellis, 1978). No significant correlation was found between the PHLC and health status ($r = -.06$). The IHLC was positively ($r = .40$, $p < .001$) and the CHLC negatively correlated ($r = -.28$, $p < .01$). Research since that time suggests that the ability of the MHLC to predict health behaviors varies by whether the respondent is healthy or has a chronic health condition. For healthy individuals, the IHLC and CHLC are better predictors than is the PHLC. The reverse is true for persons with chronic conditions, that is, the PHLC better predicts their health behaviors.

Although an empirical link has been found between poorly controlled seizures and external perceptions of control (Gehlert, 1994), external perceptions of control have been implicated in the development of psychosocial problems in epilepsy (Gehlert, 1994; Hermann, 1979; Peterson, Maier, & Seligman, 1993), and adults with epilepsy have been found to exhibit

more externality on the MHLC than did the normative sample (Gehlert, 1996; Wallston, Wallston, & DeVellis, 1978), no psychometric assessment of the MHLC in epilepsy has been done. Such an assessment is important because (a) adherence is a major problem in epilepsy (e.g., Shope, 1988); (b) external perceptions of control have been linked to medical nonadherence in most cases (DeWeerdt, Visser, Kok, & van der Veen, 1990; McLean & Pietroni, 1990; Wassem, 1991); (c) perceptions of control seem to be amenable to intervention (Baker, 1979; Felton & Biggs, 1972; Pierce, Schauble, & Farkas, 1970); and (d) the MHLC is the most widely used measure of health locus of control and seems particularly appropriate and feasible for testing perceptions of control in epilepsy patients. A psychometrically sound instrument for measuring health locus of control in epilepsy, therefore, would be useful in both research and practice. The present study was designed to address the ability of the MHLC to measure the health locus of control of adults with epilepsy by examining the factor structure of the MHLC scales when applied to persons with epilepsy. The structure underlying responses on the MHLC was evaluated using confirmatory maximum-likelihood factor analytic techniques (Jöreskog & Sörbom, 1993). The construct validity of the MHLC was further investigated using the Rasch rating scale model (Wright & Masters, 1982).

Method

Participants

The study target sample was St. Louis residents aged 18 and older who had been diagnosed with epilepsy by a licensed physician and were either members of the local epilepsy affiliate, patients at teaching hospital, or patients of a private neurologist. We chose patients from an advocacy group, a teaching clinic, and a private practice settings to maximize the variability of seizure control and socioeconomic status.

Mean age of the 143 participants was 36.5 years (SD = 11.97); 90.1% described themselves as white, 8.5% as black, and 1.4% as other. Forty-seven percent were female and 53% were male. Thirty-eight percent said that they were unemployed at the time of testing; 37.3% were employed, 11.3% were homemakers, 7% were students, and 6.3% were retired. Forty percent were single, 43.4% were married, 15.4% were separated or divorced, and 0.7% were other. The percentage frequency of types of epilepsy was 68.8 for generalized epilepsy and 31.2 for partial epilepsy.

The frequencies of categories of demographic variables was comparable to national norms for persons with epilepsy.[1] The sample was consistent with the population of persons with epilepsy as a whole with regard to gender and marital status (Lectenberg, 1984; Hauser & Hesdorffer, 1990). It approximated the population as a whole in terms of types of epilepsy experienced (Hauser & Hesdorffer, 1990). The study sample differed from the general population of persons with epilepsy in its underrepresentation of black and overrepresentation of white participants and its higher rate of unemployment (Hauser & Hesdorffer, 1990).

Procedure

Data were collected by mailed questionnaires. This method was chosen over telephone or in-person interviews in part because we believed that using either of those two methods might render persons without telephones or transportation unable to participate. Because a disproportionate number of persons with epilepsy live in poverty (Hauser & Hesdorffer, 1990), and are, therefore, less likely to have ready access to telephones or transportation, using telephone or in-person interviews might have resulted in persons of lower socioeconomic status with epilepsy being underrepresented in the study.

A second reason for mailing questionnaires to potential participants is that epilepsy remains a condition that many people chose not to divulge to others secondary to the perception that doing so would result in their being stigmatized and discriminated against (Dell, 1986). This makes it particularly difficult to access persons with the epilepsy for research. No national health survey data are available on persons with epilepsy nor is there any sort of repository of names of persons with epilepsy. Although not ideal, the research method used in the present study seemed the best way of accessing a socioeconomically-mixed sample of persons with epilepsy with varying levels of seizure control while ensuring confidentiality.

Questionnaires and materials for mailing were prepared by one of the authors. Staff members of the referral sources then mailed them to all individuals on their mailing lists who met the study's inclusionary criteria. This was done in order to ensure confidentiality. Of the 782 questionnaires mailed, 96 were returned as undeliverable, and 32 persons telephone the author using the number provided on the cover letter to inform her that the survey form could not be completed. Of the 654 remaining questionnaires, 143 were completed and returned, a 22% return rate.

Instruments

Participants completed Form A of the MHLC scales (Wallston, Wallston, & DeVellis, 1978). It is composed of three 6-item subscales reflecting the degree to which individuals attribute health outcomes to Internal Control, Powerful Others, and Chance. Items were rated on a 6-point Likert scale (from 1 = "strongly disagree" to 6 = "strongly agree").

Analysis

Means, standard deviations, and Cronbach's alpha reliability coefficients were calculated for each originally constituted scale. Intercorrelations between subscales were also obtained. This was done to allow comparison with other published psychometric analyses of the MHLC scales (e.g., Wallston et al., 1978). An interval scale of measurement was assumed, as has been the case historically with the MHLC scales, to make comparison possible.

A confirmatory factor analysis was conducted using the LISREL 8 computer program (Version 8.10, Jöreskog & Sörbom, 1993) to determine the extent to which the Wallston et al. (1978) three-factor model fit the sample of adults with epilepsy. The analysis was based on a priori specification, proposed by Wallston et al. (1978), that three factors characterized the data, and that the model investigated corresponded exactly to the expected structure. That is, each item was constrained to load on one and only one factor. This allowed us to empirically validate the factor structure of MHLC items. A Pearson correlation matrix was used rather than a matrix of polychoric correlations, because the study's small sample size and six response categories for each question obviated the use of the latter type of matrix.

To evaluate the unidimensionality and construct validity of the three pre-defined subscales in greater detail, Rasch rating scale analysis (Wright & Masters, 1982) was conducted for each subscale. The rating scale model specifies that the log odds of scoring in two adjacent categories is a function of three additive parameters: person ability, item difficulty, and step difficulty. The log odds is given by:

$$\ln [P_{nij} / P_{ni(j-1)}] = B_n - D_i - F_j,$$

in which P_{nij} is the probability of person n scoring in category j of item i, $P_{ni(j-1)}$ is the probability of person n scoring in category j-1 of item i, B_n is

the measure of person n, and D_i is the difficulty of item i, and F_j is the step difficulty of the threshold between categories j-1 and j. In the present study, F_1 is the transition from category 1 to category 2 and F_5 is the transition from category 5 to category 6. The BIGSTEPS computer program (Wright & Linacre, 1997) was used for Rasch analyses. Separate item calibrations were carried out for each subscale. Unweighted item fit mean square (MNSQ) values (expected value = 1.0) were also calculated to identify potential misfitting items, or those that indicate a lack of construct homogeneity with other items in a scale. This was done to assure scale unidimensionality. Items with MNSQ values outside the 0.8-1.2 range were identified as possible misfitting items meriting more careful examination according to Rasch models (Linacre & Wright, 1993, p. 4).

The purpose of using Rasch analysis was to determine whether items on the MHLC scales measured the same underlying construct with the sample of persons with epilepsy and to help define the MHLC subscales operationally. This is possible because the item hierarchy obtained using Rasch analysis reflects the underlying concept for each subscale as well as its qualitative meaning for study participants.

Results

Descriptive and reliability analyses

Interscale correlations, means, standard deviations, and Cronbach's alpha reliability coefficients for the MHLC scales resulting from raw score analysis are shown in Table 1. The only significant correlation between subscales was the PHLC subscale's significant and positive correlation with the CHLC subscale (.45, $p < .001$). Alpha reliabilities for the three subscales were .76, .79, and .70 for the IHLC, PHLC, and CHLC, respectively. The internal consistency coefficients appeared to be acceptable (0.70-0.79).

Confirmatory factor analysis

The chi-square goodness of fit statistic for the three-factor model with 132 degrees of freedom was 255.19 ($p < .001$). The goodness of fit (GFI) index for the model was .85. Bentler and Bonett's (1980) non-normed fit index (NNFI) and Bentler's (1990) comparative-fit index (CFI) were .78 and .81, respectively. These goodness of fit statistics represent the goodness of fit associated with a "null" model in which values over .9 indicate acceptable fit of the model to the data. Although the chi-square statistic

and fit statistics indicated a poor fit between the model and the data, the relative likelihood ratio (χ^2:*df*) was 1.93, reflecting an acceptable fit (Carmines & Mclver, 1981; Marsh & Hocevar, 1985; Wheaton, Muthen, Alwin, & Summers, 1977). Visual inspection of the Q-plot of the standardized residuals also suggested a reasonable fit, since the points fell close to the 45 degree line.

Factor loadings for each item on their a priori subscales are presented

Table 1

Intercorrelations Between Subscales, Means, Standard Deviations (SD), and Cronbach's Alpha Reliability Coefficients for the MHLC Scales (N = 143)

Subscale	1	2	3	Mean	SD	Alpha
1. Internal Health Locus of Control	--	-0.02	-.17	21.34	7.66	.76
2. Powerful Other Locus of Control		--	.45***	19.75	8.33	.79
3. Chance Locus of Control			--	22.94	7.98	.70

*** $p<.001$.

in Table 2. T-tests of factor loading coefficients for each item were significant ($p < .01$), ruling out the null hypothesis that the coefficients were equal to zero. All 18 items had significant loadings on the factors that corresponded to their a priori subscales. The item with the highest loading (.70) on the IHLC subscale was Item 13 ("If I take care of myself, I can avoid illness."). Item 8 ("When I get sick, I am to blame.") had the lowest factor loading (.35) on the IHLC subscale, calling into questions its fit with the subscale. The item with the highest factor loading (.68) on the PHLC subscale was Item 18 ("Regarding my health, I can only do what my doctor tells me to do."). Item 7 ("My family has a lot to do with my becoming sick or staying healthy.") had a factor loading on the PHLC subscale (.50) that was relatively low compared to the way that other items loaded. Item 11 ("My good health is largely a matter of good fortune.") had the highest loading (.75) on the CHLC. Item 15 ("No matter what I do, I'm likely to get sick.") was the only item to have a loading of less than .30. Its factor loading was a modest .29. The fit of this item with its subscale could, therefore, be questioned.

Table 2

Confirmatory Factor Analysis of the Multidimensional Health Locus of Control Scales: Standardized Item Factor Loadings and *t* Statistics

Subscales and items	Factor Loading	*t* *
Internal Health Locus of Control		
1. If I get sick, it is my own behavior which determines how soon I get well.	0.52	5.94
6. I am in control of my health.	0.68	8.09
8. When I get sick I am to blame.	0.35	3.83
12. The main thing which affects my health is what I myself do.	0.68	8.18
13. If I take care of myself, I can avoid illness.	0.70	8.46
17. If I take the right actions, I can stay healthy.	0.66	7.86
Powerful Other Health Locus of Control		
3. Having regular contact with my physician is the best way for me to avoid illness.	0.59	6.97
5. Whenever I don't feel well, I should consult a medically trained professional.	0.65	7.75
7. My family has a lot to do with my becoming sick or staying healthy.	0.50	5.63
10. Health professionals control my health.	0.66	7.89
14. When I recover from an illness, it's usually because other people have been taking good care of me.	0.64	7.60
18. Regarding my health, I can only do what my doctor tells me to do.	0.68	8.20
Chance Health Locus of Control		
2. No matter what I do, if I am going to get sick, I will get sick.	0.44	4.79
4. Most things that affect my health happen to me by accident.	0.57	6.40
9. Luck plays a big part in determining how soon I will recover from an illness.	0.63	7.20
11. My good health is largely a matter of good fortune.	0.75	8.84
15. No matter what I do, I'm likely to get sick.	0.29	3.08
16. If it's meant to be, I will stay healthy.	0.53	5.89

Note: * All *t* tests were significant at the $p < .001$ level, except for that of Item 15 ($p < .01$).

Rasch rating scale analysis

A Rasch residual factor analysis (Linacre, in press; Wright, 1997; Wright & Linacre, 1997) of item-response residuals was conducted to identify and evaluate the underlying structure of the 18-item MHLC scales. All six items of the IHLC were clearly identified as Factor 1 (see Table 3). Those 12 items identified as Factor 2 in the first analyses were subjected to a second Rasch factor analysis, and six PHLC items were found to cluster together (see Table 4). The six items that remained were from the CHLC. Item compositions for the three factors resulting from Rasch factor analysis were identical to those of the three subscales identified by Wallston et al. (1978).

Table 3

Results of Rasch Principal Component Analysis of Standardized Residuals Correlations (Sorted by Loading) using all 18 Items of the Multidimensional Health Locus of Control Scales

			INFIT	OUTFIT	
FACTOR	LOADING	MEASURE	MNSQ	MNSQ	ITEM
1	.72	-.28	.98	.96	IHLC 13
1	.70	-.49	.89	.92	IHLC 17
1	.69	-.23	.85	.87	IHLC 6
1	.67	-.44	1.06	1.07	IHLC 12
1	.60	-.32	1.14	1.18	IHLC 1
1	.43	.15	1.27	1.39	IHLC 8
1	-.48	.30	1.01	.98	PHLC 10
1	-.41	-.03	.99	1.13	CHLC 16
1	-.36	.71	1.25	1.20	CHLC 9
1	-.34	.31	1.05	1.01	CHLC 11
1	-.34	.02	.91	.90	PHLC 18
1	-.31	.33	1.34	1.42	CHLC 15
1	-.30	-.16	.77	.87	PHLC 5
1	-.30	-.01	.97	1.02	CHLC 2
1	-.27	-.20	.73	.73	PHLC 14
1	-.25	.14	.95	.95	CHLC 4
1	-.24	.36	.99	1.05	PHLC 7
1	-.14	-.16	.99	.98	PHLC 3

To investigate the three subscales further, separate item calibrations were conducted for each subscale. Real (i.e., not modeled) person separa-

Table 4

Results of Rasch principal Component Analysis of Standardized Residuals Correlations 9Sorted by Loading) Using 12 Items from the Powerful Others Externality (PHLC) and Chance Externality (CHLC) Subscales of the Multidimensional Health Locus of Control Scales

FACTOR	LOADING	INFIT MEASURE	OUTFIT MNSQ	OUTFIT MNSQ	ITEM
1	.64	-.33	1.08	1.18	PHLC 3
1	.55	.19	.91	.84	PHLC 10
1	.43	-.39	.77	.79	PHLC 14
1	.43	-.34	.79	.84	PHLC 5
1	.37	-.13	.92	.91	PHLC 18
1	.09	.26	1.05	1.14	PHLC 7
1	-.57	.20	1.04	.98	CHLC 11
1	-.54	.65	1.21	1.10	CHLC 9
1	-.48	.01	1.02	.97	CHLC 4
1	-.34	-.18	.94	1.05	CHLC 16
1	-.27	.22	1.39	1.44	CHLC 15
1	-.16	-.16	.99	.96	CHLC 2

tion reliabilities for the three subscales were .67, .70, and .59 for the IHLC, PHLC, and CHLC, respectively. Real item separation reliabilities were very high (.95) for all three subscales. Separate item calibrations for the subscales are summarized in Table 5, in which items are listed in ascending item-difficulty order. The item-difficulty hierarchies identified which items were harder or easier for participants to agree with. As an example, item 8 ("When I get sick I am to blame."), with an item difficulty of .73, was the most difficult item on the IHLC subscale for participants to agree with, whereas Item 17 ("If I take the right actions, I can stay healthy.") was the easiest to agree with (item difficulty = -.38).

As can be seen in Table 5, four items (items 8 and 17 of the IHLC, item 7 of the PHLC, and item 15 of the CHLC) were identified as possibly misfitting according to the MNSQ > 1.2 or < 0.8 rule specified. These four items operated somewhat differently than other items on their own scales. Confirmatory and Rasch factor analyses yielded similar results in situations in which items loaded relatively low on their scales.

Table 5

Rasch Analysis of Multidimensional Health Locus of Control Scales (Items Listed in Descending Order of Difficulty in Each Subscale)

Subscale, item number and content		Difficulty	Outfit MNSQ
Internal			
8.	When I get sick I am to blame.	.73	1.72
6.	I am in control of my health.	.06	.84
13.	If I take care of myself, I can avoid illness.	-.02	.88
1.	If I get sick, it is my own behavior which determines how soon I get well.	-.09	1.12
12.	The main thing which affects my health is what I myself do.	-.30	.85
17.	If I take the right actions, I can stay healthy.	-.38	.76
Powerful others			
14.	When I recover from an illness, it's usually because other people have been taking good care of me.	.36	.94
5.	Whenever I don't feel well, I should consult a medically trained professional.	.29	.94
3.	Having regular contact with my physician is the best way for me to avoid illness.	.29	1.04
18.	Regarding my health, I can only do what my doctor tells me to me.	.01	1.00
10.	Health professionals control my health.	-.44	.83
7.	My family has a lot to do with my becoming sick or staying healthy.	-.52	1.49
Chance			
16.	If it's meant to be, I will stay healthy.	.34	.87
2.	No matter what I do, if I am going to get sick, I will get sick.	.31	1.01
4.	Most things that affect my health happen to me by accident.	.12	.96
11.	My good health is largely a matter of good fortune.	-.09	.82
15.	No matter what I do, I'm likely to get sick.	-.11	1.28
9.	Luck plays a big part in determining how soon I will recover from an illness.	-.58	.89

Note: Item difficulties for each scale were from separate analyses. A higher number on item difficulty indicates that it was more difficult to disagree with the item.

Discussion

In evaluating the study results, the possible effect of the low response rate of (22%) must be considered. Persons with epilepsy are an extremely difficult group to access and recruit for research. Consequently, little is known about either their health (not even accurate prevalence data are available) or psychosocial status. Although the profile of study subjects was very similar to national norms for persons with epilepsy, results should be considered exploratory and might have been different had a greater number of subjects participated in the study.

The results of the present study suggest that the Multidimensional Health Locus of Control scales possess reliable and valid psychometric properties when used with adults with epilepsy. Descriptive statistics for each subscale were similar to those reported for other samples (Winefield, 1982; Wallston, Wallston, & DeVellis, 1978; Hartke & Kunce, 1982; Winefield, 1982; Umlauf & Frank, 1986). The instrument had an acceptable level of internal consistency (.70-.79). Cronbach's alpha coefficients are generally similar to those reported by Wallston et al. (1978) for the MHLC derivation sample and were higher than those in Winefield's (1982) sample of medical and dental students.

Confirmatory and Rasch factor analyses lent support to the hypothesized three-factor structure of the measure. Evaluation of the structure underlying responses to the scales demonstrated reasonable construct validity for each of the three subscales. The three MHLC subscales seem to measure separate dimensions of locus of control beliefs related to health, namely, internal, powerful others, and chance, with possible exceptions. Epilepsy is known to seriously disrupt family relationships and interactions (Lectenberg, 1984). It is, therefore, not surprising that the one item (Item 7, "My family has a lot to do with my becoming sick or staying healthy") on the MHLC subscales that focuses on family influence on health behavior would misfit. Similarly, that Item 8 ("When I get sick I am to blame") misfit is not surprising, considering the growing body of empirical evidence on the attributional style of persons with epilepsy which shows a strong tendency to blame oneself for failure and attribute success to others (Gehlert, 1996).

The MHLC appears to have the same factor structure when used with adults with epilepsy as reported for the derivation sample, as evidenced by confirmatory and Rasch analyses. Results of the two approaches were

generally congruent and complementary. The Rasch approach shed additional light on the construct by positioning items on a continuum.

That the MHLC scales demonstrated psychometric soundness in the present study suggests that it would be a useful tool for measuring health locus of control in adults with epilepsy. This has implications for both practice and research. Adherence to medical regimens, primarily taking anticonvulsant medications, is a major problem in epilepsy treatment (e.g., Shope, 1988). Although external perceptions of control have not yet been linked empirically to nonadherence in epilepsy, they have in other patient groups (e.g., Wassem, 1991). Because perceptions of control are malleable via psychotherapy, it may be the case that identifying externality and providing appropriate intervention would be a means of increasing medical adherence in persons with epilepsy. The present study also suggests that the MHLC scales can be used with confidence in measuring locus of control for purposes of research. Locus of control has received recent attention among researchers as an etiological factor in psychosocial problems in epilepsy (Gehlert, 1996; von Steinbuchel, Krauth, Scheidereiter, & Hiltbrunner, 1996). Having a valid and reliable tool for measuring the construct would facilitate this area of research.

Footnote

[1] The gender ratio of the present study was 1:1.1, compared to 1.2:1 in the population of persons with epilepsy as a whole (Hopkins, 1987). Lectenberg (1984) reports that 56% of males and 69% of females with epilepsy marry. Fifty-five percent of males and 64% of females in the present study were either married at the time that they completed the questionnaire, or said that they were separated, divorced, or widowed. Hauser and Hesdorffer (1990) state that 39-59% of seizures experienced have a generalized onset and that 32-52% are partial. In the present study, 66% of participants reported having generalized seizures and 30% reported having partial epilepsy. Ninety-one percent of respondents reported that they were White, 8% said that they were Black, and 0.7% reported that they were either American Indian or "other." Hauser and Hesdorffer (1990) report a 1.3-2.2 times greater incidence of epilepsy among Black males than among White males, and a 1.4-1.7 times greater incidence among Black than among White females. Thirty-nine percent of the study sample reported that they were unemployed. The federal unemployment rate was 6.4% in the month and year in which data were gathered. Thus, the study

sample does not reflect the Epilepsy Foundation of America's estimate of an unemployment rate in epilepsy that is over twice that of the population as a whole (Ann Scherer, personal communication, October 25, 1997).

References

Baker, E.K. (1979). The relationship between locus of control and psychotherapy: A review of the literature. *Psychotherapy: Theory, Research and Practice, 16,* 351-362.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Burish, T. G., Carey, M. P., Wallston, K. A., Stein, M. J., Jamison, R. N., & Lyles, J. N. (1984). Health locus of control and chronic disease: An external orientation may be advantageous. *Journal of Social and Clinical Psychology, 2,* 326-332.

Carmines, E. G., & Mclver, J. P. (1981). Analyzing models with unobservable variables: Analysis of covariance structures. In G. W. Bohrnstedt & E. F. Borgatta (Eds.), *Social measurement: Current issues* (pp. 65-115). Beverly Hill, CA: Sage.

Coelho, R. J. (1985). A psychometric investigation of the Multidimensional Health Locus of Control Scale with cigarette smokers. *Journal of Clinical Psychology, 41,* 372-376.

Dell, J.L. (1986). Social dimensions of epilepsy: Stigma and response. In S. Whitman and B.P. Hermann (Eds.), *Psychopathology in epilepsy: Social dimensions* (pp. 185-210). New York: Oxford University Press.

DeVellis, R. F., DeVellis, B. M., Wallston, B. S., & Wallston, K. A. (1980). Epilepsy and learned helplessness. *Basic and Applied Social Psychology, 1,* 241-253.

DeWeerdt, I., Visser, A.P., Kok, G., & van der Veen, E.A. (1990). Determinants of active self-care behavior of insulin treated patients with diabetes: Implications for diabetes education. *Social Science and Medicine, 40,* 605-615.

Felton, G.S., & Biggs, B.E. (1972). Teaching internalization behavior to collegiate low achievers in group psychotherapy. *Psychotherapy: Theory, Research, & Practice, 9,* 281-283.

Galanos, A.N., Strauss, R.P., & Pieper, C.F. (1994). Sociodemographic correlates of health beliefs among black and white community dwelling elderly individuals. *Internal Journal of Aging and Human Development, 38,* 339-350.

Gehlert, S. (1994). Perceptions of control in adults with epilepsy. *Epilepsia, 35,* 81-88.

Gehlert, S. (1996). Attributional style and locus of control in adults with epilepsy. *Journal of Health Psychology, 1,* 469-477.

Hartke, R. J., & Kunce, J. T. (1982). Multidimensionality of health-related locus-of-control-scale items. *Journal of Consulting and Clinical Psychology, 50,* 594-595.

Hauser, W.A., & Hesdorffer, D.C. (1990). *Epilepsy: Frequency, causes, and consequences.* New York: Demos.

Hatz, P. S. (1978). *The relationship of life satisfaction and locus of control in patients undergoing chronic hemodialysis.* Unpublished master's thesis. University of Illinois, Champaign.

Hermann, B.P. (1979). Psychopathology in epilepsy and learned helplessness. *Medical Hypotheses, 5,* 723-729.

Hopkins, A. (1987). Definitions and epidemiology of epilepsy. In A. Hopkins (Ed.), *Epilepsy* (pp. 1-17). New York: Demos.

Horlick, L., Cameron, R., Firor, W., Bhalerao, U., & Baltzan, R. (1984). The effects of education and group discussion in the post myocardial infraction patient. *Journal of Psychosomatic Research, 28,* 485-492.

Ingle, R. J., Burish, T. G., & Wallston, K. A. (1984). Conditionability of cancer therapy patients. *Oncology Nursing Forum, 11(4),* 97-102.

Jöreskog, K.G., & Sörbom, D. (1993). LISREL 8.10 and PRELIS 2.10 for Windows [Computer software]. Chicago: Scientific Software International, Inc.

Kucera-Bozarth, K., Beck, M.S., & Lyss, L. (1982). Compliance with lithium regimens. *Journal of Psychosocial Nursing and Mental Sciences, 29,* 11-15.

Lectenberg, R. (1984). *Epilepsy and the family.* Cambridge: Harvard University Press.

Levenson, H. (1973). Multidimensional locus of control in psychiatric patients. *Journal of Consulting and Clinical Psychology, 41,* 397-404.

Levenson, H. (1974). Activism and powerful others: Distinctions within the concept of internal-external control. *Journal of Personality Assessment, 38,* 377-383.

Levenson, H. (1975). Multidimensional locus of control in prison inmates. *Journal of Applied Social Psychology, 5,* 342-347.

Linacre, J.M. (in press). Factor analysis of residuals. *Journal of Outcome Measurement.*

Linacre, J. M., & Wright, B. D. (1993). *A user's guide to BIGSTEPS.* Chicago: MESA Press.

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept. *Psychological Bulletin, 97,* 562-582.

McCallum, D.M., Keith, B.R., & Wiebe, D.J. (1988). Comparison of response formats for Multidimensional Health Locus of Control scales: Six levels versus two levels. *Journal of Personality Assessment, 52,* 732-736.

McLean, J., & Pietroni, P. (1990). Self care: Who does best? *Social Science and Medicine, 30,* 591-596.

Nagelberg, D.B. (1979). *Evaluating the BGSU health risk reduction program: A comparison of differing methods of providing health information to college students.* Unpublished doctoral dissertation, Bowling Green State University, Bowling Green, Ohio.

Nagy, V. T., & Wolfe, G. R. (1983). Chronic illness and health locus of control beliefs. *Journal of Social and Clinical Psychology, 1,* 58-65.

O'Looney, B. A., & Barrett, P. T. (1983). A psychometric investigation of the Multidimensional Health Locus of Control questionnaire. *Journal of Clinical Psychology, 22,* 217-218.

Peterson, C., Maier, S.F., & Seligman, M.E.P. (1993). *Learned helplessness: A theory for the age of personal control.* New York: Oxford University Press.

Pierce, R.M., Schauble, P.G., & Farkas, A. (1970). Teaching internalization behavior to clients. *Psychotherapy: Theory, Research, & Practice, 7,* 217-220.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Robinson-Whelen S. & Storandt, M. (1992). Factorial structure of two health belief measures among older adults. *Psychology and Aging, 7(2),* 209-213.

Rock, D. L., Meyerowitz, B. E., Maisto, S. A., & Wallston, K. A. (1987). The derivation and validation of six multidimensional health locus of control scale clusters. *Research in Nursing and Health, 10,* 185-195.

Rotter, J. (1966). Generalized expectations for internal versus external control of reinforcement. *Psychological Monographs, 80,* 1-28.

Russell, S. F., & Ludenia, K. (1983). The psychometric properties of the multidimensional health locus of control scales in an alcoholic population. *Journal of Clinical Psychology, 39,* 453-459.

Sherwin, D. (1979). *Self-care health practices and beliefs pertaining to hypertensive young adults.* Unpublished doctoral dissertation, University of Wisconsin at Milwaukee.

Shipley, R.H. (1981). Maintenance of smoking cessation: Effect of follow-up letters, smoking motivation, muscle tension, and health locus of control. *Journal of Consulting and Clinical Psychology, 49,* 982-984.

Shope, J.T. (1988). Compliance in children and adults: Review of studies. In D. Schmidt & I.E. Leppik (Eds.), *Compliance in epilepsy* (pp. 23-47). Amsterdam: Elsevier.

Umlauf, R. L., & Frank, R. G. (1986). Multidimensional health locus of control in a rehabilitation setting. *Journal of Clinical Psychology, 42,* 126-128.

von Steinbuchel, N., Krauth, S., Scheidereiter, U., & Hiltbrunner, B. (1996). Role of life events, coping, social support, and health locus of control for the quality of life for people with epilepsy [Abstract]. *Epilepsia, 37* (suppl. 4), 9.

Wall, R. E., Hinrichsen, G. A., & Pollack, S. (1989). Psychometric characteristics of the Multidimensional Health Locus of Control Scales among psychiatric patients. *Journal of Clinical Psychology, 45(1),* 94-98.

Wallston, K.A. (1989). Assessment of control in health-care settings. In A. Stepoe & A. Appels (Eds.), *Stress, personal control, and health* (pp. 85-105). Chichester, England: Wiley.

Wallston, K.A., Stein, M.J., & Smith, C.A. (1994). Form C of the MHLC scales: A condition-specific measure of locus of control. *Journal of Personality Assessment, 63,* 534-553.

Wallston, K. A., Wallston, B. S., & DeVellis, R. (1978). Development of the Multidimensional Health Locus of Control (MHLC) scales. *Health Education Monographs, 6,* 160-170.

Wallston, K., Wallston, B., Kaplan, E., & Maides, S. (1976). Development and validation of the Health Locus of Control (HLC) scale. *Journal of Consulting and Clinical Psychology, 44,* 580-585.

Wassem, R. (1991). A test of the relationship between health locus of control and the course of multiple sclerosis. *Rehabilitation Nursing, 16,* 189-93.

Wheaton, B., Muthen, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. In D. R. Heise (Ed.), *Sociological methodology* (pp. 84-136). San Francisco: Jossey-Bass.

Winefield, H. R. (1982). Reliability and validity of the health locus of control scale. *Journal of Personality Assessment, 46,* 614-619.

Wright, B. D. (1997). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling, 3*(1), 3-24.

Wright, B. D., & Linacre, J. M. (1997). *BIGSTEPS: Rasch analysis for all two-facet models (Version 2.77) [Computer software].* Chicago: MESA Press.

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

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

JOURNAL OF OUTCOME MEASUREMENT®, 2(3), 191 - 221

Analyzing Nonadditive Conjoint Structures: Compounding Events by Rasch Model Probabilities

George Karabatsos
Department of Biometry and Genetics
Lousiana State University Medical Center

The following study proposes a Rasch method to measure variables of nonadditive conjoint structures, where dichotomous response combinations are evaluated. In this framework, *both* the number of endorsed items and their latent positions are considered. This is different from the cumulative response process (measurable by the Rasch model), where the probability of a positive response to an item with measure δ_i is considered a monotonic increasing function of the person's measure β_v. This is also unlike the unfolding framework, where the probability of a positive response is maximum when $\beta_v = \delta_i$, and monotonically decreases as $|\beta_v - \delta_i|$ approaches infinity.

The method involves four steps. In Step 1, items are scaled by the Rasch model for paired comparisons to produce a variable definition. These scale values serve as a basis for Steps 2 and 4. In Step 2, the nonadditive conjoint system is restructured to additive. The quantitative hypothesis of the restructured data is tested by the axioms of conjoint measurement theory in Step 3. This data is then analyzed by the Rasch rating scale model in Step 4 to evaluate individual response combinations, using the Step 1 item calibrations as anchors.

The method was applied to simulated person responses of the Schedule of Recent Events (Holmes and Rahe, 1967). The results suggest that the method is useful and effective. It scales items with a robust method of paired comparisons, ensures additivity and quantification of the conjoint person-item matrix, produces a reasonable ordering of person measures from the perspective of individual response combinations, and provides satisfactory person and item separation (i.e., reliability). Furthermore, the restructured data reproduces SRE item scale values obtained by paired comparisons in Step 1.

Requests for reprints should be sent to George Karabatsos, MESA Psychometric Laboratory, The University of Chicago, 5835 S. Kimbark Ave., Chicago, IL 60637-1609.

The cumulative response process is most frequently found in latent-trait measurement, and was implicit in the pioneering work of Binet and Simon (1905) and Thurstone (e.g., 1925,1926). Letting β_v represent the measure of person v, and δ_i represent the measure of item i, this process involves the person responding positively (x = 1) to items located at or below their position on a unidimensional variable ($\beta_v > \delta_i$), and negatively (x = 0) to items above that point ($\beta_v < \delta_i$). Such a process defines a cumulative variable definition. In a math test, items arrange from easy addition, to subtraction, to multiplication and to hard division. In attitude measurement, items arrange from less to more extreme statements.

Louis Guttman (1944) posited that person responses to items should form a deterministic cumulative process:

$$\text{Prob}\{x_{vi} = 1\} = \begin{cases} 0 \text{ if } \beta_v < \delta_i \\ 1 \text{ if } \beta_v > \delta_i \end{cases} \quad (1)$$

The scalogram (2) illustrates Guttman's expectation. 5 individual responses to 5 items are listed in increasing order of δ:

		Item (2)
		12345
	A	11111
	B	11110
Person	C	11100
	D	11000
	E	10000

It can be inferred from (2) that $\beta_A > \beta_B > \beta_C > \beta_D > \beta_E$ and $\delta_5 > \delta_4 > \delta_3 > \delta_2 > \delta_1$. In this framework, the total score indicates exactly which items have or have not been endorsed (Guttman, 1944, p. 144). Items and persons which violate this strict conjoint order are considered unreliable, and subsequently selected out of the analysis. The flaw in this scaling technique is its inability to handle observation error, which renders it impractical for most work. Although Guttman attempted to deal with this issue with a coefficient of reproducibility, it was not entirely successful (see Torgerson, 1958, p. 322-324).

Georg Rasch (1960) posed a practical alternative by specifying the cumulative response process to be stochastic. The Rasch model is expressed as:

$$\text{Prob}[x_{vi} = 1 \mid \beta_v, \delta_i] \equiv (e^{\beta v - \delta i}) / (1 + e^{\beta v - \delta i}), \quad (3)$$

where β_v is a linear function of the logit proportion of positive responses by person v across all i, and δ_i is a linear function of the logit proportion of negative responses contained in item i across all v, centered at zero (Wright and Stone, 1979). An attractive property of the model is that person and item scores are sufficient statistics for the parameters β_v and δ_i. Figure 1 shows the cumulative response function of the Rasch model, where the probability of a positive response for person v on item i depends on the logit distance between β_v and δ_i:

$$\text{Prob}\{x_{vi} = 1\} = \begin{cases} \{1 < p > .50 & \text{if } \beta_v - \delta_i > 0 \\ \{.50 & \text{if } \beta_v - \delta_i = 0 \\ \{0 < p < .50 & \text{if } \beta_v - \delta_i < 0 . \end{cases} \tag{4}$$

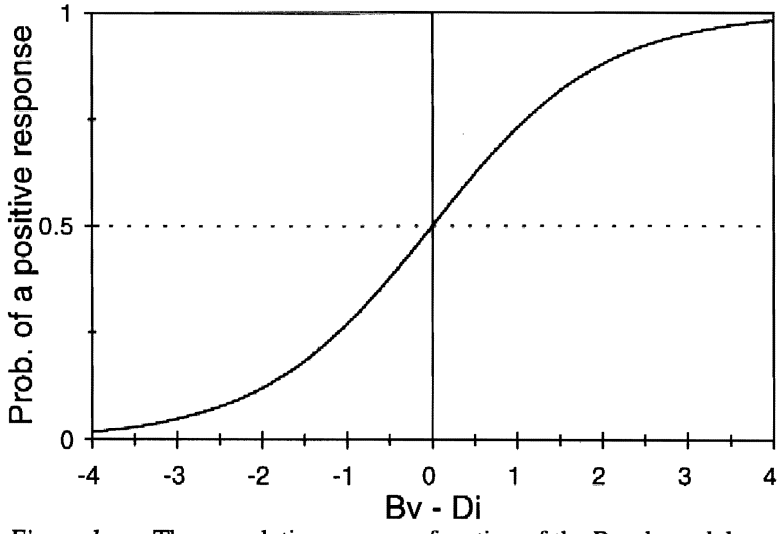

Figure 1. The cumulative response function of the Rasch model.

The stochastic interpretation of item responses prevents, instead of allows, minor response deviations from disturbing person measurement. This is why the model is useful in a wide range of measurement applications.

$$\begin{array}{c} \mathbf{1110101000} \\ \uparrow \\ \beta_v \end{array} \tag{5a}$$

(5a) illustrates a response string which fits Rasch model specifications. Here, the person endorses the first three "easy" items ($\beta_v - \delta_i > 0$), followed by a stochastic combination of 1's and 0's near the person's measure ($\beta_v - \delta_i \approx 0$), concluding with a consistent string of 0's ($\beta_v - \delta_i < 0$). Verification of such a stochastic structure in data is provided by two mean square (MNSQ) statistics. Infit MNSQ is sensitive to unexpected responses made by persons when the distance between β_v and δ_i is small. Outfit MNSQ detects unexpected person responses when this distance is large. Mean square values less than .7 identify an overly predictable, Guttman structure. Values ranging from .7 to 1.3 identify measurable stochasticity (i.e., (4)) and local independence. A stochastic response pattern is preferable to a Guttman pattern. This is because Guttman patterns, when split in two parts, produces an easy test on which respondents performed infinitely well, and a difficult test on which they performed infinitely poorly (Linacre and Wright, 1994). MNSQ > 1.3 identify improbable responses, as unpredictable persons (or items) are immeasurable in the cumulative framework. (5b) is a response string with unpredictable responses, which provides no cumulative evidence of the person's position on the variable.

0101010101 (5b)

↑ ↑ ↑ ↑ ↑

β_v?

Items and persons which misfit the model are usually removed from the analysis. Such data distorts the transitivity among items and persons necessary for an additive conjoint data structure.

Additive Conjoint Measurement

N. R. Campbell (1920) deduced that fundamental measurement requires additive properties, and therefore its construction is to be performed through concatenation. It is easy to see how fundamental measurement is obtained in physical science, where concatenations of length and weight are explicit. The combined length of several rods is determined by joining them end-to-end. The combined weight of several bricks is determined by piling them on top of one other. On the other hand, concatenation operations are rare in the social sciences. A person's measure on a particular attitude cannot be concatenated by the summation of ordinal Likert responses. If measurement is to be constructed from ordinal variables, the data structure needs to approximate concatenation.

Luce and Tukey (1964) deduced additive conjoint measurement as the only means of providing quantitative structure to ordinal data:

> The essential character of what is classically considered, e.g., by N.R. Campbell, the fundamental measurement of extensive quantities is described by an axiomatization for the comparison of effects of (or responses to) arbitrary combinations of "quantities" of a *single specified kind* ... Measurement on interval scales which have a common unit follows from these axioms; usually these scales can be converted in a natural way into ratio scales. (p. 1)
>
> A close relation exists between conjoint measurement and the establishment of response measures in a two-way table, or other analysis-of-variance situations, for which "the effects of columns" and "the effects of rows" are additive. Indeed, the discovery of such measures ... may be viewed as the discovery, via conjoint measurement, of fundamental measures of the row and column variables. (p. 1)
>
> In...the behavioral and biological sciences, where factors producing orderable effects and responses deserves more useful and more fundamental measurement, the moral seems clear: When no natural concatenation operation exists, one should try to discover a way to measure factors and responses such that the "effects" of different factors are additive. (p. 4)

Conjoint measurement is described in several sources (e.g., Luce and Tukey, 1964; Coombs, Dawes, and Tversky 1970, Ch. 2; Krantz, Luce, et al., 1971; Narens, 1985; Michell, 1990; 1988). The following description follows Michell.

Additive conjoint measurement involves a variable, P, to be a noninteractive function of two other variables, I and J. This function may either be additive ($P = f(I + J)$) or multiplicative ($P = f(I * J)$). This is analogous to a multi-level ANOVA design, where the independent variables I and J are ordinal, and produce non-interactive significant effects on the dependent variable P. I, J, and P form a conjoint system when: (1) P can have an infinite number of values, (2) $P = f(I, J)$, (3) there is a weak order ($\succeq$) among the values of P, and (4) that values of I and J are identifiable. The conjoint system is quantitative when it satisfies solvability, the Archimedian condition, independence, and double cancellation.

Solvability is satisfied when every value of P occurs within every row I and every column J of the conjoint matrix, i.e., for any a, b ∈ I and x ∈ J, a value y ∈ J exists, such that ax = by. The Archimedian condition holds when any two values of I, J, or P are never infinitely larger than any other two values of I, J, or P. Within a finite conjoint system, the Archimedian and solvability conditions are not directly testable. However, they are indirectly confirmed when independence and double cancellation are satisfied.

Eight double cancellation tests have been used in research. When independence (or single cancellation) is satisfied, only Luce-Tukey double cancellation needs to be tested, since it is the only version which is falsi-

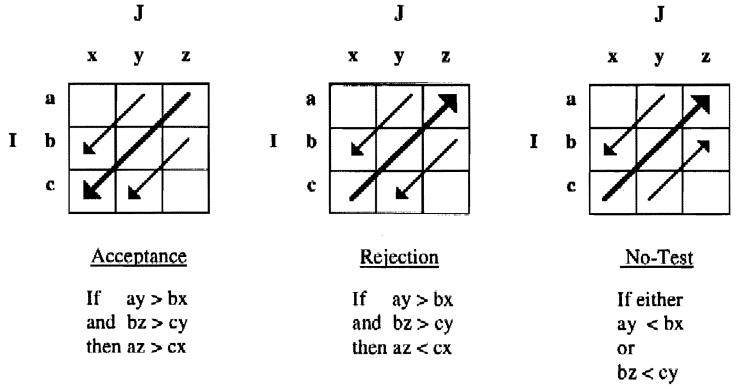

Figure 2. The possible outcomes when testing the Luce-Tukey double cancellation axiom.

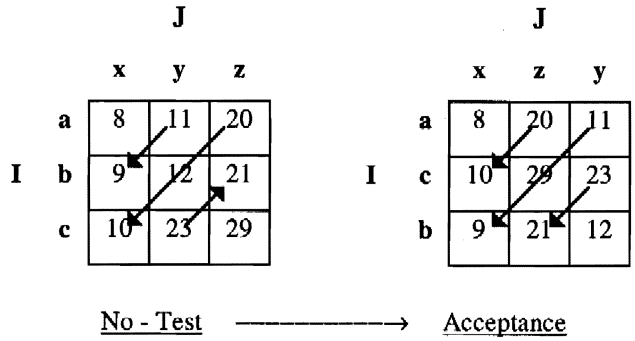

Figure 3. A demonstration of how the "No-Test" outcome is transformed into "Acceptance."

fiable upon confirmation of independence. Independence is met when the values of P arrange in weak order ($\succeq$), increasing from left to right within every row, and from top to bottom within every column.

There are three possible outcomes for the test of double cancellation, as shown in Figure 2. For any a, b, c $\in$ I and x, y, z $\in$ J in a 3 X 3 submatrice of a conjoint system, the Luce-Tukey double cancellation axiom is satisfied when the antecedents ay $\geq$ bx and bz $\geq$ cy and the consequent az $\geq$ cx are true. Double cancellation is violated when the antecedents are true and the consequent is false. The "No Test" outcome occurs when at least one of the antecedents is false. It has been proven, however, that by rearranging elements within I and/or J, "No-Tests" can be transformed into "Acceptance" outcomes. Since the double cancellation condition considers a, b, and c to be *any* three values of I, and x, y, z to be *any* three values of J, there are 36 (3! * 3!) substitution instances within a 3 X 3 matrix. Figure 3 demonstrates a case with hypothetical data: by reordering the rows from abc to acb, and the columns from xyz to xzy, the outcome is transformed.

Despite the inherent power of conjoint measurement, its application in social science has been limited. Cliff (1992) suggests that the causes are abstract mathematics and the lack of examples to which social scientists can relate. Most important, since errors inevitably occur in observation, Cliff concludes that there doesn't appear to be many ways to deal with axiom failure. Hence, the virtue of conjoint measurement is found when applied in a practical manner.

There is a definite connection between Rasch and additive conjoint measurement. In the Rasch framework, the conjoint system is conceptualized as $P = f(\beta + \varepsilon)$ or $P = f(\beta * \varepsilon)$, where P is the proportion of positive responses, b is the person measure, and e is the inverse of the item measure δ. The multifaceted Rasch model (Linacre, 1989/94) is expressed as $P = f(\beta + \varepsilon + L)$, where L represents judge leniency. Data fit to the Rasch model implies that the axioms of additive conjoint measurement are satisfied, and that items and persons are measured on a common interval scale (Brogden, 1977). Furthermore, this outcome is analogous to person and item statistical sufficiency (Fisher, 1922), infinitely divisible parameters (Levy, 1924; Kolmogorov, 1950), and parameter separation (Rasch, 1960), all necessary conditions for sample and test-free measurement (see Wright, 1997).

Perline, Wright, and Wainer (1979) report that, using data which fit

the model, 93% of 3 X 3 submatrices satisfied the Luce-Tukey test of double cancellation, while 83% satisfied with less fitting data (treating "No-Test" as "Acceptance"). In the latter case, item fit was reasonable (mean outfit MNSQ = .91), indicative of the highly structured data which conjoint measurement axioms demand. Since the probabilistic specifications of the Rasch model prevent slightly deviating responses from disturbing the interpretation of the data structure, it is considered a practical form of conjoint measurement, as person and item fit statistics are more reasonable and informative alternatives to testing double cancellation. On the other hand, the violation of double cancellation, whether minimal or extreme, is evidence against a quantitative data structure.

Nonadditive Conjoint Structures

Hence, data containing cumulative responses which fit Rasch specifications manifest conjoint additivity. Recall that in a cumulative response process, a positive response to one item implies a positive response to any less extreme item. A success on a division item implies success on addition, subtraction, and multiplication items. An endorsement of an extreme attitude implies that less extreme statements are endorsed. However, there are other item types, namely "point items", which do not facilitate cumulative response processes. Torgerson (1958) explains the difference between cumulative and point items:

> ...with the point items, a positive response means 'I am here,' whereas, with the corresponding monotone items [which facilitate cumulative responses] the positive response means 'I am above this,' and a negative response means 'I am below this.' (p. 312)

Hence, with point (or nonmonotone) items, items (δ_i) and persons (β_v) are not adequately represented by logit-proportional frequency counts.

There are two ways in which point item responses can be interpreted. The first is through methods of unfolding, where only the location of positive responses are of interest. In stochastic unfolding, the probability of a positive response is at a maximum when $\beta_v - \delta_i = 0$, and monotonically decreases as $\beta_v - \delta_i$ approaches $\pm \infty$. Unfolding analysis can be seen with the following application: Imagine a six-item political attitudes survey, where the first two reflect liberal beliefs, the third and fourth reflect moderate beliefs, and the last two reflect a conservative attitude. Possible response strings of a liberal would be: 110000, 101000, 111000, a moder-

ate: 011000, 001100, 000110, 001010, and a conservative: 000011, 000101, 000111. Examples of response strings that do not fit the unfolding framework are 110011, 100010, and 010001. Since unfolding is not the primary focus of this study, the theory will not be discussed further. Readers interested in unfolding may refer to Coombs (1964), Andrich (1997), and Hoijtink (1997). Linacre (1993) provides an interesting approach by restructuring folded data to fit the cumulative framework of the Rasch model.

A second way in which to analyze point items is to evaluate the combinations of positive responses, where *both* the number of endorsed items and their latent positions are considered. In this framework, response strings which misfit in cumulative and unfolding models necessarily occur. Such applications arise when the scientist is confronted with a nonadditive data structure, and therefore is compelled to measure event combinations.

Table 1 is an example of the second scenario, which shows the item statistics of a five item crime survey administered to 576 people (52.3% criminals, 44.7% non-criminals, 3% unclassified).[1] The survey assesses

Table 1
Rasch Scale Values of Crime Survey

RAW SCORE	MEASURE	ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS CORR.	CRIME
50	1.29	.17	1.03	.3	1.38	2.0	-.10	Vagrancy
56	1.12	.16	1.33	3.3	1.85	4.4	-.34	Homicide
128	-.38	.13	.72	-5.5	.68	-5.2	.21	Receive Stolen Goods
161	-.97	.13	1.11	2.1	1.06	.8	-.18	Assault & Battery
166	-1.06	.13	.77	-4.6	.75	-3.3	.12	Larceny
MEAN	.00	.15	.99	-.9	1.15	-.3		
S.D.	1.01	.02	.22	3.5	.43	3.5		

Note: This table shows the Rasch scale values of the crime survey's five items which was administered to 576 criminals and non-criminals. The item hierarchy is based on the frequency of positive responses (raw score). Observe that items do not arrange in severity order.

each person's level of criminality by asking to indicate the committed offenses. Although items (mean infit= .98, mean outfit= 1.14) and persons (mean infit = .99, mean outfit = 1.15) fit the Rasch model, the negative point-biserial correlations is evidence that noncumulative person response strings occurred frequently in the data, which causes the item hierarchy to misrepresent the crime severity variable. This Rasch analysis defines the crime variable to be, in the order of most to least severe, Vagrancy (1.29) > Homicide (1.12) > Receiving Stolen Goods (-.38) > Assault and Battery (-.97) > Larceny (-1.06). On the other hand, Thurstone (1927) obtained a more plausible hierarchy through his method of paired-comparisons: Homicide (3.16) > Assault and Battery (1.47) > Larceny (1.33) > Receiving Stolen Goods (1.00) > Vagrancy (0.00).

When the Thurstone scale values were applied to the data, fit to the Rasch model decreased (mean item infit MNSQ = 1.38, Mean item outfit MNSQ = 1.80). Table 2 gives the statistics of all possible response strings, where items are arranged in severity order. The person measures, as based on the number of positive responses, concludes that 10011 > (00101 = 01001 = 11000) > (00001 = 10000), where 10011, 00101, 01001, and 00001 misfit the Rasch model. Unfolding response models would conclude that 00001 > 00101 > 10011 > 01001 > 11000 > 10000, where 01001, 10011, and 01001 misfit. Similar examples are found throughout Table 2.

However, in terms of criminal severity, one would expect 10011 > 00101 > 01001 > 00001 > 11000 > 10000. Is there a measurement system that can produce such a result, where the combinations of crimes are evaluated? Furthermore, how can misfitting responses be included within psychometric analysis, yet retain the conjoint additivity necessary for fundamental measurement?

M.H. Birnbaum has performed considerable research in the study of combination judgments. His most recent application (Birnbaum and Sotoodeh, 1991) involved the Schedule of Recent Events (SRE, Holmes and Rahe, 1967), a questionnaire which asks respondents to indicate stressful events that have occurred during a given time period (e.g., past six months). The SRE intends to measure the life stress of an individual by interpreting the combination of events. Using techniques of mathematical psychology, he made the following conclusions (Birnbaum and Sotoodeh, 1991, italics added):

Table 2
Person Measure and Fit Statistics of 28 Different Response Strings

PERSON NUMBER	TOTAL SCORE	PERSON MEASURE	ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS CORR.	RESPONSE STRING
1	5	4.01	1.58	MAXIMUM ESTIMATED MEASURE					11111
2	4	3.06	1.23	1.57	.7	1.43	.3	-.69	11011
3	4	3.06	1.23	1.89	1.0	4.59	1.8	.67	01111
4	4	3.06	1.23	1.69	.8	1.88	.6	-.13	10111
5	4	3.06	1.23	.44	-1.0	.29	-.9	-.13	11110
6	3	1.85	1.02	2.08	1.9	2.57	2.0	-.11	01011
7	3	1.85	1.02	1.74	1.4	1.71	1.1	-.57	10011
8	3	1.85	1.02	2.23	2.1	2.74	2.2	.56	00111
9	3	1.85	1.02	2.01	1.8	2.51	1.9	-.11	01101
10	3	1.85	1.02	.89	-.3	.81	-.4	-.11	10110
11	3	1.85	1.02	.66	-.9	.58	-.9	-.57	11100
12	3	1.85	1.02	1.22	.5	1.67	1.0	.56	01110
13	2	.87	1.00	2.03	2.3	3.05	2.2	.00	00101
14	2	.87	1.00	1.87	2.0	2.91	2.1	-.49	01001
15	2	.87	1.00	2.03	2.3	3.05	2.2	.00	00101
16	2	.87	1.00	.65	-1.1	.57	-.8	-.86	11000
17	2	.87	1.00	1.35	.9	1.22	.3	.00	01010
18	2	.87	1.00	2.09	2.4	3.12	2.3	.00	00011
19	2	.87	1.00	.88	-.4	.77	-.4	-.49	10010
20	2	.87	1.00	.81	-.6	.71	-.5	-.49	10100
21	2	.87	1.00	1.51	1.3	1.35	.5	.75	00110
22	2	.87	1.00	1.28	.7	1.15	.2	.00	01100
23	1	-.27	1.18	1.71	.9	6.44	2.1	-.38	00001
24	1	-.27	1.18	1.39	.5	1.40	.3	.30	00010
25	1	-.27	1.18	.59	-.8	.40	-.6	-.87	10000
26	1	-.27	1.18	1.19	.3	.95	.0	-.38	01000
27	1	-.27	1.18	1.33	.5	1.24	.2	.30	00100
28	0	-1.16	1.54	MINIMUM ESTIMATED MEASURE					00000

Note: This table shows the person measure and fit statistics of the 28 different response strings found in the data for the 5-item crime survey based on the number of positive responses (total score). These results use the crime scale values of Thurstone (1927).

> Judgments of "combinations" of stressful events were not simply the sums of their separate events; instead, *they showed two systematic departures from additivity*. First, the effect of a given event was less than when it was the least stressful than when it was the most, as if the most stressful event carries extra configural weight. Second, each additional stressor has a diminishing marginal effect on the overall judgment. (p. 236) ... *the data show subadditivity* (p. 242).
>
> When one stressful event is included, the overall judgment is high, and the other [stress] events have less effect. *This convergence is consistent with the idea that the most stressful event carries the greatest weight in each combination* (p. 240).

These observations mirror findings of studies where people were judged according to their moral and immoral deeds (Birnbaum, 1973; Riskey and Birnbaum, 1974), credibility (Birnbaum, Wong, and Wong, 1976), and their likableness as expressed by adjectives (Birnbaum, 1974):

> Impressions of likableness cannot be represented as simple sums or averages of single values of the adjectives. They appear to be a predictable, but nonadditive function of the component values. The data shows consistent, regular deviations from additivity that are similar for different selections of the adjectives. When one adjective is dislikable, the person is rated as dislikable, and variation of the other trait has less affect. *Differential or configural weighting of the more dislikable traits can account for the interactions*....Representation of the stimuli by distributions could explain why the adjectives are integrated by a nonadditive function: It is less likely for a person possessing a dislikable trait to be likable than for a person with a likable trait to be dislikable. (Birnbaum, 1974, p. 560, italics added)

He further argues against the use of conjoint measurement when measuring combinations:

> Conjoint measurement analysis (Krantz and Tversky, 1971) describes conditions that ideal data would have to satisfy to be *ordinally* consistent with the theory. For example, crossover interactions [when measuring the combinations of positive responses] would be ordinally inconsistent with additive models, for no monotonic transformation could ever make

> the data fit the model. In this case, everyone agrees that the model should be rejected. It should be noted that ordinal violations of additivity will show up as significant interactions in the analysis of variance. The problem arises when significant interactions occur in the *absence* of ordinal violations. (Birnbaum, 1974, p. 545)

In mathematical psychology, increased attention has recently been placed in the measurement of nonadditive representations (e.g., Luce, Krantz, Suppes, and Tversky, 1990). As Narens and Luce (1993) observe:

> These scale type ideas [laws of combination] play an increasing role in achieving suitable psychophysical laws (Falmagne, 1985) and in offering families of nonadditive representations...*examples arise in the analysis of nonadditive conjoint structures*. (p. 128-129, italics added)

Hence, while Birnbaum argues against the use of conjoint measurement for combination judgments, Narens and Luce suggest that the measurement of combinations is possible within this framework.

To scale the magnitude of items, particularly those of the SRE, Birnbaum recommends the convergence of several different models to produce a "unified" scale of stressful events (Birnbaum and Sotoodeh, 1991, p. 241):

$$r_{ij} = s_j - s_i \tag{6}$$

$$d_{ij} = s_j - s_i \tag{7}$$

$$C_i = w_1 (s_i) + b_C \tag{8}$$

$$C_{ij} = w_2 (s_i + s_j + \omega | s_i - s_j |) + b_C \tag{9}$$

$$C_{ijk} = w_3 (s_i + s_j + s_k + \omega | s_{max} - s_{min} |) + b_C \tag{10}$$

(6) is used to scale stressful events based on the respondents' ratio judgments between event pairs s_j and s_i (e.g., 1/8. 1/4, ½, 1, 2, 4, 8). (7) is used for "difference" judgments between s_j and s_i pairings, using a 200 point rating scale where 0 = No stress difference between the pair, 100 = event 1 is very much more stressful than event 2, and -100 = event 2 is very much more stressful than event 1. (6) and (7) have the same form, since judgments of "ratios" and "differences" are often monotonically related (Birnbaum, 1978, 1980, 1982; Birnbaum and Jou 1990). This monotonic relationship was confirmed in his SRE analysis. (8), (9), and (10) are the configural weighting models for the combination judgments

of one, two, and three events, respectively. These are also based on rating scale judgments, where 0 = the combination is not stressful at all, and 100 = the combination is maximally stressful. For configural weighing, the weight of each event depends on its rank among other occurred events. ω is the configural weight taken from s_{min} and transferred to s_{max}, or vice versa, depending on its result. The coefficients w_1, w_2, w_3, and b_C are constants, and evidence of an additive data structure is indicated when $w_1 = w_2 = w_3$ and $\omega = 0$. In all, Birnbaum had 95 subjects scale 15 stressful events, which required nineteen parameters (p. 241). His research design resulted in 309 judgments per subject.

Despite the apparent utility of Birnbaum's method, there are issues related to its practical use. Are hundreds of judgments per subject, and such large rating scales, needed to scale items? Furthermore, are arbitrarily complex algorithms (6-10) required to measure event combinations? It is argued that more parsimonious models, specifically the Rasch model for paired comparisons, and the Rasch rating scale model, provide a more useful approach to combination judgments, once additivity is recovered in the conjoint structure. The method to be proposed will recover additivity by using the Rasch model's stochastic specifications.

Method

Instrument. Ten dichotomous items of the SRE will illustrate the proposed method. They include "Death of Spouse," "Jail Term," "Divorce," "Fired at work," "Marriage," "Child leaving home," "Moving," "New family member," "Christmas," and "Change in family get-togethers." The SRE is known as a measure of life change, and has been used extensively to predict patients' onset of psychosomatic illness and psychiatric disorders.

Studies of life-event instruments like the SRE have been criticized for low internal consistency reliability (e.g., Katschnig, 1986; Steele, et al., 1980; Lei and Skinner, 1979; Brown, 1974; Mendels and Weinstein, 1972). However, these findings have been misunderstood. Since stress events are not necessarily correlated with each other, the SRE should have zero reliability (Clearly, 1981). The reliability coefficient is a function of inter-item correlations, and inter-item correlations are lower for point than cumulative items (Torgerson, 1958, p.317).

Subjects. To simulate a noncumulative data structure, data will consist of response strings with all possible combinations of one, two, or

three positive responses among the ten SRE items (N=175), reflecting the occurrence of stressful events during the past six months. The idea is that no more than three stressful events can possibly occur for each person during this time period.

Plan of analysis. The objective is to recover additivity for the conjoint structure, and subsequently interpret person measures, which evaluates stress event combinations. There are four steps to this process:

(1) Item scaling. The inclusion of misfitting response strings within Rasch analysis distorts the item hierarchy. Unless the hierarchy produced by simultaneous person-item Rasch calibrations is plausible (after removing misfitting data), it is recommended that paired-comparisons be used. The Rasch model for paired-comparisons, for dichotomous choices without ties, specifies the probability that stimulus m is chosen over stimulus n (m > n) is given by:

$$\text{Prob}\,[m > n \mid \beta_m, \beta_n] \equiv (e^{\beta m - \beta n}) / (1 + e^{\beta m - \beta n}), \qquad (11)$$

Observe that (11) follows the additive form of conjoint measurement, and that the absence of a person parameter minimizes sample-dependent estimates. (11) is the same as the Bradley-Terry-Luce model (Luce, 1959), and the "ratio" and "difference" models ((6) and (7)). However, the Rasch version is most practical, since it is a probabilistic rather than a descriptive model (Linacre, 1989/94). Therefore, it is able to handle missing data, and detect unpredictable choices.

To scale the SRE items, thirty graduate students of the University of Chicago were asked to choose the most stressful event (without ties) among every possible pairings of the 10 SRE events, producing 45 judgments per respondent. The SRE item scale generated by this step will serve as a basis for data restructuring (Step 2) and person measurement (Step 4). Furthermore, this scale will be compared to the one obtained by Birnbaum's complex algorithms (6-10). If they are similar, it follows that (6-10) are unnecessary.

(2) Data restructuring. To restructure the data, the probability of a positive response to each item will be reconceptualized from $\beta_v - \delta_i$ to $\delta_{MAX} - \delta_i$, where δ_{MAX} represents the calibration of the maximum item endorsed by person v. The probability values produced by $\delta_{MAX} - \delta_i$ will be multiplied by 10 to produce a 0 - 10 rating scale, replacing the 0/1 responses in the original data. Positive responses to items with calibrations less than δ_{MAX} ($\delta_i < \delta_{MAX}$) are assigned a rating of 10. To reflect the

characteristics of combination judgments described by Birnbaum's research, the data restructuring method assigns higher scores to persons with high δ_{MAX} values than low δ_{MAX} values, and specifies greater weight for positive responses to items near δ_{MAX} than those further away.

(3) Verifying conjoint additivity of the restructured data. The independence and the Luce-Tukey double cancellation axioms will test whether the restructured conjoint data is quantitative. Independence will be verified empirically with Kendall's T_C, a nonparametric statistic which correlates the rank order of the data to a specified criterion order (Kendall, 1970). In this application, two correlations will be computed. The first will use the criterion of strict order ($\succ$) within all rows, where the values of P increase from left to right. The second will use the criterion of strict order within all columns, where P increases from top to bottom. Hence, positive ($T_C > 0$) correlations within rows and columns verifies the weak order ($\succeq$) requirement of independence ($z > 1.65$, α=.05, one-tailed test). Double cancellation will be tested by calculating the percentage of all 3 X 3 submatrices in the conjoint system that produce an "Acceptance" outcome.

Mean-square statistics were considered for testing data fit to the additive hypothesis. However, considering that the scoring method must reduce the stochasticity of the data structure, these statistics will diagnose overfit. Therefore, it is appropriate to pose sharper tests to the data.

(4) Person Measurement. The newly formed rating scale data will be analyzed by the Rasch model of ordered response categories:

$$\text{Prob}\,[x_{vi} = \kappa \mid \beta_v, \delta_i, T_\kappa] \equiv (e^{\beta v - (\delta i + T\kappa)}) / (1 + e^{\beta v - (\delta i + T\kappa)}) \quad (12)$$

where T_κ (centered at zero) represents the step measure of rating scale category κ from $\kappa - 1$ (Wright and Masters, 1982). The SRE scale values obtained in paired comparisons (Step 1) will serve as anchors for person measurement.

The transitivity among person measures will be verified by observing the combinations of stress events that occurred for each individual. For instance, suppose that person A had 2 very stressful events, person B experienced 1 very stressful event, and Person C encountered 2 mild events and 1 moderate one. The person measurement order among these three cases should be A > B > C. Hence, the proposed method can be perceived as a weighted-measurement scheme.

Separation statistics will estimate the extent to which persons and items identify a useful variable line. Separation (SEP) is a ratio equal to the square root of the true variance (TV) divided by error variance (EV), or SEP = $(TV/EV)^{½}$. The relationship between separation and reliability (REL) is REL = $(SEP)^2 / (1 + SEP)^2$. Person separation describes the number of performance levels that the test measures among the sample of respondents, while item separation indicates how well items spread along the variable.

To determine whether the measured sample is discriminated by 11 categories, the 0 - 10 rating scale will be analyzed from three perspectives. The first examines the average measure of persons responding to a particular rating scale category. Ideally, every advancement in the rating scale corresponds with an increase in the average person measure. The second perspective requires the step measure (T_k) to increase with every rating scale progression. The third perspective investigates the infit and outfit statistics of each category.

Finally, a Rasch analysis will be performed on the restructured data *without item anchoring*. Ideally, this analysis will produce an SRE item scale similar to the one obtained by Rasch paired-comparisons in Step 1. Such a result demonstrates that the unanchored Rasch analysis produces similar person measures as the analysis with item anchors.

Results

Item Scaling. Table 3 summarizes the hierarchy of the 10 SRE stressful events (from least to most stressful), as computed by the Rasch paired-comparisons model. A striking result is that the Pearson correlation between the Rasch and Birnbaum SRE scales is .97, despite the fact that Birnbaum's models are more complex, involve 3 types of elaborate rating scales (e.g., 200-point rating scale), and used three times the number of subjects.

Many of the pairs, such as "Christmas-Death of Spouse" and "Jail term-Moving," elicit predictable responses. Furthermore, the Rasch model is known to produce stable calibrations, even with small sample sizes (see Lord, 1980). These two ideas suggest that the entire 45 X 30 paired-comparison matrix is not required to produce a useful scale of stressful events. Figure 4 demonstrates that after removing 17 people from the data, and retaining only pairs within ± 1 logit of each other (as indicated in Table 3), the Birnbaum and Rasch scales remain the same (r = .97). "Jail

Table 3
Hierarchy of Stressful Events

STRESS Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Events
3.04	.23	0.4	-1	1.0	0	10 Death of spouse
2.52	.21	1.3	0	1.9	1	9 Jail term
2.21	.21	0.9	0	1.6	1	8 Divorce
.80	.19	1.2	0	1.0	0	7 Fired at work
-.62	.17	0.9	0	1.2	0	6 Marriage
-.76	.17	0.7	0	0.7	0	5 Child leaving home
-.87	.17	1.1	0	1.8	1	4 Moving
-1.20	.17	1.0	0	1.0	0	3 New family member
-2.34	.20	1.1	0	1.3	0	2 Christmas
-2.78	.22	1.8	1	2.1	1	1 Change/family get togethers
.00	.19	1.1	0.0	1.3	0.5	Mean
1.94	.02	0.4	0.8	0.4	0.7	S.D.

Note: This table shows hierarchy of SRE stressful events, as calibrated by the Rasch dichotomous paired comparisons model, where 30 respondents each made 45 paired-comparison judgements.

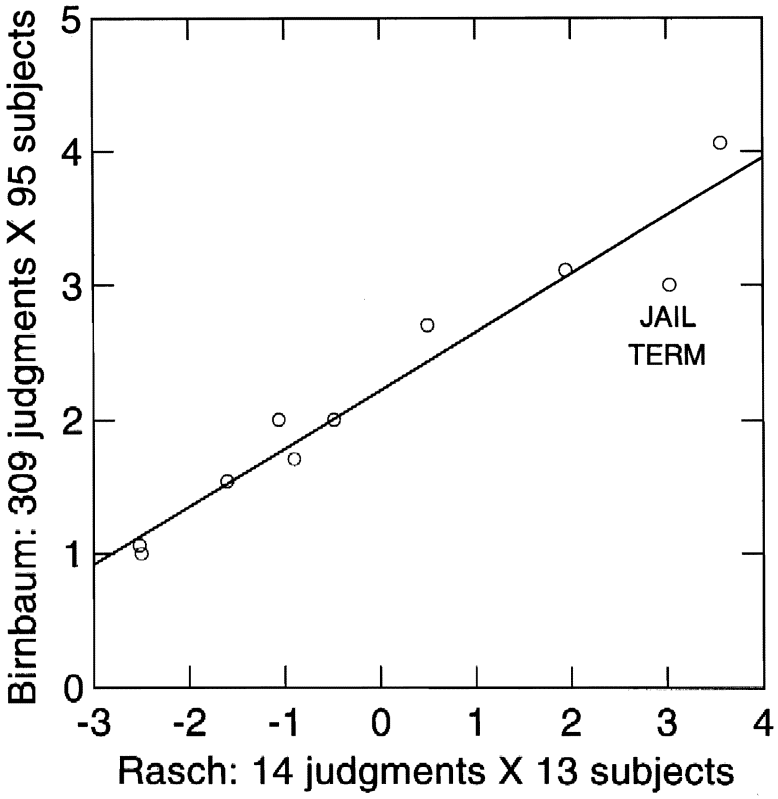

Figure 4. A comparison of Birnbaum and Rasch scale values. Note that the Rasch method used 95% fewer judgments and 86% fewer subjects, but yielded a similar scale (r = .97).

term" is the only disagreement, of which it can be argued that the Rasch model scales more plausibly. Hence, comparable results were obtained between the two methods, even though the Rasch model analyzed 86% fewer subjects, and used 95% fewer judgments per subject.

Data restructuring. Table 4 presents the person scoring system which restructures the dichotomous data, among individuals whose maximum item endorsed is 10, 9, 8, or 7. Recall that δ_{MAX} pertains to the measure of the maximum item endorsed. Therefore, if a person's most extreme item is 8, for example, $\delta_{MAX} = 2.21$.

Also, the Table shows that the probability of a positive response is attained by calculating

$$\text{Prob}[x_{vi} = 1 \mid \delta_{MAX}, \delta_i] \equiv (e^{\delta max - \delta i}) / (1 + e^{\delta max - \delta i}). \qquad (13)$$

Table 4
Data Restructuring When Maximum Item Endorsed is 10, 9, 8, or 7

	Stress Event (δ_i)									
	1 (-2.78)	2 (-2.34)	3 (-1.20)	4 (-.87)	5 (-.76)	6 (-.62)	7 (.80)	8 (2.21)	9 (2.52)	10 (3.04)
$\delta_{MAX} = 3.04$										
Response String	0	0	0	0	0	0	0	0	0	1
$Px_{vi} = \frac{(e^{\delta max - \delta_i})}{(1 + e^{\delta max - \delta_i})}$	.99	.99	.98	.98	.98	.97	.90	.70	.63	.50
X 10 = Rating	10	10	10	10	10	10	9	7	6	5
$\delta_{MAX} = 2.52$										
Response String	0	0	0	0	0	0	0	0	1	0
$Px_{vi} = \frac{(e^{\delta max - \delta_i})}{(1 + e^{\delta max - \delta_i})}$	.99	.99	.98	.97	.96	.96	.85	.58	.50	.37
X 10 = Rating	10	10	10	10	10	10	9	6	5	4
$\delta_{MAX} = 2.21$										
Response String	0	0	0	0	0	0	0	1	0	0
$Px_{vi} = \frac{(e^{\delta max - \delta_i})}{(1 + e^{\delta max - \delta_i})}$	.99	.99	.97	.96	.95	.95	.80	.50	.42	.30
X 10 = Rating	10	10	10	10	10	10	8	5	4	3
$\delta_{MAX} = 0.80$										
Response String	0	0	0	0	0	0	1	0	0	0
$Px_{vi} = \frac{(e^{\delta max - \delta_i})}{(1 + e^{\delta max - \delta_i})}$	.97	.96	.88	.84	.83	.80	.50	.20	.15	.10
X 10 = Rating	10	10	9	8	8	8	5	2	2	1

Note: This table illustrates data restructing when the maximum item endorsed is either 10, 9, 8, or 7. Positive responses to items less extreme than δ_{MAX} are assigned a rating of 10.

The obtained probability value is multiplied by 10 to produce the 0 - 10 rating scale. The rating scale values attained across all items constitutes as the new person response string. Suppose that a person's original response string is 0010101000, in which the maximum item endorsed is "Fired at Work" (item 7, $\delta_{MAX} = 0.80$). Initially, that individual's response string will become 10 10 9 8 8 8 5 2 2 1. Since there are also positive responses to items 3 and 5, each will automatically be assigned a rating of "10," producing the final response string of 10 10 10 8 10 8 5 2 2 1.

Notice that a higher weight was assigned for a positive response to item 5 than item 3, as 10 - 8 = 2 for item 5, and 10 - 9 = 1 for item 3. This is because $\delta_{MAX} - \delta_5 < \delta_{MAX} - \delta_3$, since $\delta_5 > \delta_3$. For a given δ_{MAX}, positive responses to items further away has less effect on the overall score than positive responses to items near δ_{MAX}. Therefore, the occurrence of "Moving" adds more stress to a person who has a "Child leaving home" than to a person who faces a "Jail term," an event which is already very stressful. Using crime as an example, a vagrancy offense would make a thief seem as a worse criminal. But vagrancy does not add much to the severe criminality status of the murderer.

Verifying conjoint additivity of the restructured data. Table 5 displays the 9 X 4 conjoint system, where the stress score proportion (P) is a function of person stress (β) and item stress (ε).[2] By glancing at the Table, one can see that the data structure is ideal: The stress proportions are greater for high stress versus low stress person groups, and greater for severe versus mild stress events.

Table 5
9 x 4 Conjoint System of Restructured SRE Data

	Rounded Person Score Group	(Severe) ← Stress Events → (Mild) 8 - 10	6 - 7	3 - 5	1 - 2	Number of persons with each score
(Low)	10	.00	.05	.13	.45	1
	20	.00	.10	.20	.65	2
↑	30	.00	.25	.43	.80	1
Person	40	.01	.31	.55	.92	17
Stress	50	.02	.35	.72	.92	20
↓	60	.17	.65	.85	1.00	7
	70	.17	.69	.89	1.00	15
	80	.44	.94	1.00	1.00	51
(High)	90	.63	.96	1.00	1.00	61

The 9 x 4 conjoint system of the restructured SRE data 9n=175). The dependent variable represents proportion of the rating scale endorsed.

Table 6
Results of Independence and Double Cancellation for the Restructured SRE Data

Independence	Tc	Z
Within rows	1.00	4.48
Within columns	.92	4.83
Double Cancellation	**N**	**%**
3 X 3 submatrices satisfy	330	98
3 X 3 submatrices violate	6	2
Total tests	336	

Table 6 reports that the data satisfies independence ($p < .001$). Furthermore, the conjoint system contains 336 distinct 3 X 3 submatrices ((9!/6!3!) * (4!/1!3!)), of which 98% satisfy double cancellation. These results indicate that the restructured data is quantitative, confirming that items and persons are measurable on a common interval (or ratio) scale.

Person Measurement. Table 7 displays the person measures of 43 out of the 175 simulated individuals, using the SRE item calibrations (Table 3) as anchors. The statistics of the entire sample are available from the author. It is clear that person measures are a function of δ_{MAX}. However, δ_{MAX} does not entirely determine person measure order, as person 47 > 18, 148 > 143, 153 > 147, 161 > 149, and 163 > 152. The fact that 47 > 46 suggests the combination of "Jail term," "Divorce," and "Fired at work" is more stressful than "Death of Spouse." Person 152 = 158 because "Marriage" is as stressful as the combination of "Change in family get-togethers," "Christmas," and "Child leaving home." On the other hand, 134 > 136 means that "Fired at work" is more stressful than the combination of "New family member," "Child leaving home," and "Marriage."

The analysis of the original 0/1 data matrix indicate very low person (SEP=.00, REL=.00) and item separation (SEP=1.28, REL=.62), as person (mean infit MNSQ= 1.88, mean outfit MNSQ= 5.87) and item fit (mean infit MNSQ= 4.05, mean outfit MNSQ= 4.10) is unsatisfactory. On the other hand, the person and item separation of the restructured data is 4.59 (REL=.95) and 20.49 (REL=1.0), respectively, indicating that the SRE items are more on target with the measured sample. As expected with this data, there is overfit among persons (mean infit MNSQ=.41,

Table 7
A Comparison of 43 Person Measures
After Data Restructured

PERSON NUMBER	TOTAL SCORE	PERSON MEASURE	ERROR	RESPONSE STRING
1	94	4.07	.37	0000000111
2	92	3.81	.34	0000001011
10	91	3.70	.34	0000000011
11	90	3.58	.34	0000010101
47	89	3.47	.34	0000001110
18	88	3.35	.35	0000011001
46	87	3.22	.36	0000000001
60	85	2.95	.37	1000001010
83	84	2.81	.38	0000000010
84	82	2.53	.36	0000101100
91	80	2.29	.34	0000110100
115	67	1.07	.29	0001101000
117	66	.99	.28	0010101000
126	65	.92	.27	0000101000
127	64	.84	.27	0110001000
134	63	.77	.26	0000001000
136	52	.11	.24	0010110000
137	51	.05	.25	0011010000
138	49	-.07	.25	0100110000
148	49	-.07	.25	0011100000
143	48	-.14	.26	0000110000
144	48	-.14	.26	1010010000
145	47	-.20	.26	0001010000
151	47	-.20	.26	0101100000
153	46	-.27	.27	1001100000
147	45	-.35	.27	1100010000
155	45	-.35	.27	1010100000
161	45	-.35	.27	0111000000
149	44	-.42	.28	0100010000
157	44	-.42	.28	0010100000
163	44	-.42	.28	1011000000
152	43	-.50	.28	0000010000
158	43	-.50	.28	1100100000
164	43	-.50	.28	0011000000
159	42	-.58	.28	0100100000
165	42	-.58	.28	1101000000
166	41	-.67	.29	0101000000
167	40	-.75	.28	1001000000
168	39	-.83	.28	0001000000
169	38	-.91	.28	1110000000
170	36	-1.05	.27	0110000000
172	34	-1.20	.26	0010000000
173	23	-1.94	.28	1100000000

mean outfit MNSQ=.27) and items (mean infit MNSQ=.57, mean outfit MNSQ=.27).

The rating scale analysis of the restructured data is shown in Table 8. Due to the nature of the rescoring technique, the "Average Measure" results are ideal. Category 10 indicates misfit, probably due to the "10" ratings assigned for positive responses to items when $\delta_i < \delta_{MAX}$. Furthermore, the step measures are out-of-order, as they do not increase with every advancement of the rating scale. These results suggest that a more parsimonious rating scale may be used without decreasing person measurement quality.

Table 8
Rating Scale Analysis of the 0 -10 Categorization

CATEGORY LABEL	OBSERVED COUNT	AVGE MEASURE	INFIT MNSQ	OUTFIT MNSQ	STEP MEASURE
0	110	-3.25	.24	.49	NONE
1	45	-2.19	.04	.04	-1.68
2	87	-1.31	.13	.08	-2.42
3	29	-.69	.09	.07	.15
4	86	-.14	.17	.20	-1.37
5	211	.31	.15	.14	-.75
6	100	.74	.15	.09	1.21
7	37	1.14	.18	.08	1.78
8	85	1.62	.18	.08	.38
9	115	2.36	.19	.13	1.57
10	845	3.94	2.02	1.52	1.12

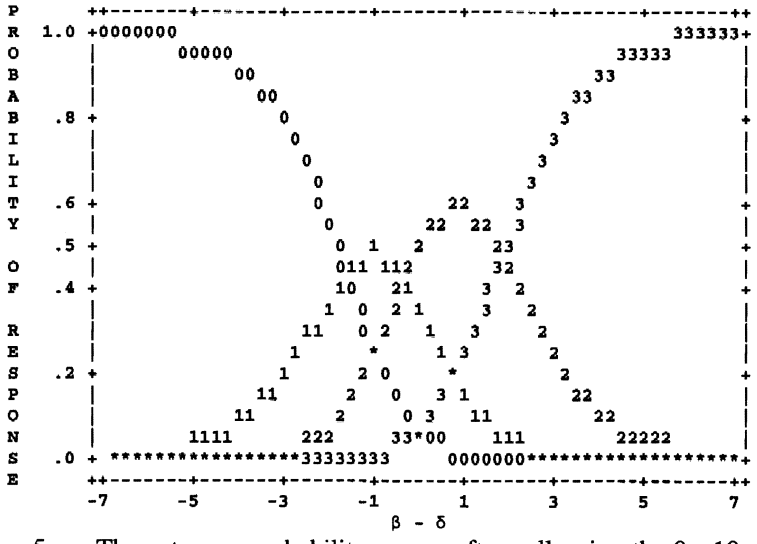

Figure 5. The category probability curves after collapsing the 0 - 10 rating scale into 0 - 3.

Therefore, the next Rasch analysis collapsed the rating scale into four categories, from 0 1 2 3 4 5 6 7 8 9 10 to 0 1 1 1 2 2 2 2 3 3 3. This change produced ideal category probability curves (see Figure 5), where each category curve at some point on the (β - δ) continuum is higher than the other three, in an ordered fashion. Figure 6 compares person measurements produced by the 11 category versus the 4 category

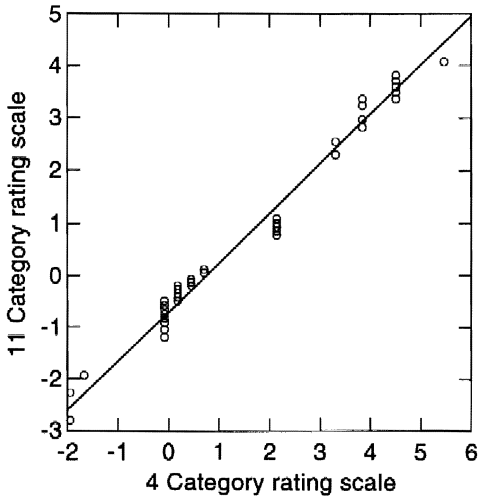

Figure 6. A comparison of person measures using different rating scales: 11 categories (0-10) versus 4 categories (0-3). The Pearson correlation is .99.

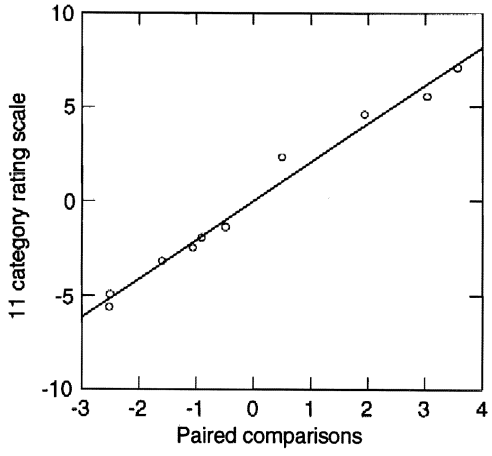

Figure 7. A comparison of SRE item calibrations using different methods: 11 category Rasch analysis versus Rasch paired comparisons (r = .99)

rating scale, and shows that they are the same (r = .99). Hence, the SRE data may be analyzed with the 0 - 3 rating scale without losing information in person measurement.

To determine whether the restructured data reproduces the SRE item calibrations obtained in Step 1, the 0-10 rating-scale data was Rasch analyzed with the SRE items *unanchored*. Figure 7 shows that this analysis produces an SRE item scale identical to the one obtained by paired comparisons in Step 1 (r =.99). The unanchored Rasch analysis of the 0 - 3 rating scale data also produces a scale similar to paired-comparisons (r = .99). Hence, one may analyze the restructured data without item anchors, and obtain person measures identical to those generated with item anchors. It is suggested, however, that item anchors be used in future applications. They enable Rasch programs to converge quickly in parameter estimations, as unanchored Rasch analysis of highly deterministic data require many iterations for convergence.

Discussion

Rasch techniques are well established for measuring variables of additive representations, but this work attempts to bridge the gap between Rasch models and variables of nonadditive representations. As Luce, et al.(1990) state:

> Most work in the theory of measurement has been based on additive representations for various kinds of structures. There are good reasons, however, for also studying structures in which the numerical combination rules are intrinsically nonadditive....We need to distinguish three types of numerical representations for conjoint structures: additive representations, nonadditive representations that can be transformed to additive, and representations that are essentially nonadditive (p. 18).

This study illustrates the second case, as nonadditive representations were transformed (through data restructuring) to additive. Although the method presented does not simultaneously measure items and persons, it is still convenient because items are scaled with minimum effort with the Rasch paired-comparisons model, persons are scored with a simple computer program, and measured via rating scale analysis using Rasch software (e.g., Linacre and Wright, 1998; Linacre, 1998). Furthermore, the Rasch rating scale model and Rasch paired comparisons are more practical and elegant than the models presented in (6-10). (6-10) are deterministic,

cannot handle missing data, and do not detect observation errors.

The validity of the proposed method is demonstrated in five ways. One, this method scales items with a robust paired comparisons method. Two, it constructs interval scaled measurement by ensuring conjoint additivity. Three, it produces a reasonable ordering of person measures from the perspective of event combinations. Four, compared to the original 0/1 data matrix, Rasch analysis of the restructured data indicates improved person and item separation. And five, the restructured data reproduces the item scale obtained by paired-comparisons in Step 1.

The measurement of event combinations may be applied to a wide variety of settings. Medical and psychiatric research, in particular, utilize many instruments which serve as symptom checklists (e.g., "PILL Symptom Checklist," Pennebaker, 1980). Often, symptoms do not occur cumulatively, and therefore may not produce a plausible item hierarchy within the additive framework. In such cases, it is suggested that a panel of 10 to 20 experts complete a paired-comparison questionnaire to produce a useful variable definition of symptom severity. The person-item data matrix would then be restructured to recover conjoint additivity, so that patients can be measured according to symptom combinations.

Footnotes

[1] The crime data was collected and provided by Raymond Knight and associates of Brandeis University.

[2] Person scores were rounded off, and items were grouped according to their calibrations. This was done to fit within the constraints of the conjoint analysis computer program, which only handles up to 10 X 4 conjoint systems.

References

Andrich, D. (1997). A hyperbolic cosine IRT model for unfolding direct responses of persons to items. In W.S. van der Linden and R.K. Hambleton [Eds.], *Handbook of Modern Item Response Theory* (pp. 399-414). New York: Springer-Verlag.

Binet, A. and Simon, T. (1905). MÁthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'AnnÁe Psychologique*, 11, 191-244.

Birnbaum, M.H. (1972). Morality judgments: Test of an averaging model with differential weights. *Journal of Experimental Psychology*, 93, 35-42.

Birnbaum, M.H. (1974). The nonadditivity of personality impressions. *Journal of Experimental Psychology*, 102, 543-561.

Birnbaum, M.H. (1978). Differences and ratios in psychological measurement. In N.J. Castellan and F. Restle (Eds.), *Cognitive Theory* (p. 33-74). Hillsdale, NJ: Lawrence Earlbaum Associates.

Birnbaum, M.H. (1980). Comparison of two theories of "ratio" and "difference" judgments. *Journal of Experimental Psychology: General*, 109, 304-319.

Birnbaum, M.H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social Attitudes and Psychophysical Measurement* (pp. 401-485). Hillsdale, NJ: Lawrence Earlbaum Associates.

Birnbaum, M.H., and Jou, J.W. (1990). A theory of comparative response times and "difference" judgments. *Cognitive Psychology*, 22, 184-210.

Birnbaum, M.H., and Sotoodeh, Y. (1991). Measurement of stress: Scaling the magnitudes of life changes. *Psychological Science*, 2, 236-243.

Birnbaum, M.H., Wong, R., and Wong, L.K. (1976). Combining information from sources that vary in credibility. *Memory and Cognition*, 4, 330-336.

Brogden, H.E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika*, 42, 631-634.

Brown, G.W. (1974). Meaning, measurement, and stress of life events. In B.S. Dohrenwend and B.P. Dohrenwend (Eds.) *Stressful life events: Their nature and effects* (pp. 217-243). New York: Wiley.

Campbell, N.R. (1920). *Physics: The elements.* Cambridge University Press.

Clearly, P.J. (1981). Problems of internal consistency in life-event schedules. *Journal of Psychosomatic Research*, 25, 309-320.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, 3, 186-190.

Coombs, C.H. (1964). *A theory of data.* New York: Wiley.

Coombs, C.H., Dawes, R.M., and Tversky, A. (1970). *Mathematical psychology: An elementary introduction.* Englewood Cliffs, NJ: Prentice Hall.

Falmagne, J.C. (1985). *Elements of psychophysical theory.* New York: Oxford University Press.

Fisher, R.A. (1922). On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society of London (A)*, 222: 309-368.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139-150.

Hoijtink, H. (1997). PARELLA: An IRT model for parallelogram analysis. In W.S. van der Linden and R.K. Hambleton [Eds.], *Handbook of Modern Item Response Theory* (pp. 415-429) New York: Springer-Verlag.

Holmes, J.H., and Rahe, R.H. (1967). The social readjustment rating scale. *Journal of Psychosomatic Research*, 11, 213-218.

Karabatsos, G. (1997). Sexual Experiences Survey: Interpretation and validity. *Journal of Outcome Measurement*, 1, 305-328.

Karabatsos, G. (1998). Occam's Razor at work. *Transactions of the Rasch Measurement SIG: American Educational Research Association*, 11, 587-588.

Katschnig, H. (1986). Measuring life stress: A comparison of the checklist and panel technique.

In H. Katschnig (Eds.) *Life events and psychiatric disorders: Controversial issues*, (pp. 74-106). Cambridge: Cambridge University.

Keats, J.A. (1967). Test theory. *Annual review of psychology*, 18, 217-238.

Kendall, M.G. (1970). *Rank correlation methods* (second edition). London: Griffin.

Kolmogorov, A.N. (1950). *Foundations of the theory of probability.* New York: Chelsea.

Krantz, D.H., Tversky, A. (1971). Conjoint measurement analysis of composition rules in psychology. *Psychological Review*, 78, 151-169.

Krantz, D.H., Luce, R.D., Suppes, P., and Tversky, A. (1971). *Foundations of measurement, Volume I: Additive and polynomial representations.* New York: Academic.

Lei, H., and Skinner, H.A. (1980). A psychometric study of life events and social readjustment. *Journal of Psychosomatic Research*, 24, 57-65.

Levy, P. (1937). *Theorie de l'addition des variables aleatoires* [Combination theory of unpredictable variables]. Paris: Wiley.

Linacre, J.M. (1993). *An approach to unfolding*. A paper presented at the International Objective Measurement Workshop: Atlanta, Georgia.

Linacre, J.M. (1989/94). *Many facet Rasch measurement.* Chicago: MESA.

Linacre, J.M. (1998). A user's guide to FACETS: Rasch model computer program. Chicago: MESA.

Linacre, J.M., and Wright, B.D. (1994). Chi-square fit statistics. *Transactions of the Rasch Measurement SIG: American Educational Research Association*, 8, 360-361.

Linacre, J.M., and Wright, B.D. (1998). A user's guide to BIGSTEPS: Rasch model computer program. Chicago: MESA.

Lord, F.M. (1980). Small N justifies Rasch methods. In D.J. Weiss [Ed.], Proceedings of the 1979 computer adaptive testing conference. Minneapolis: The University of Minnesota.

Luce, R.D. (1959). *Individual choice behavior.* New York: Wiley.

Luce, R.D., and Tukey, J.W. (1964). Simultaneous conjoint measurement: A new kind of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1-27.

Luce, R.D., Krantz, D.H., Suppes, P., and Tversky, A. (1990). *Foundations of measurement, Volume III: Representation, axiomatization, and invariance.* San Diego: Academic.

Mendels, J., and Weinstein, N. (1972). The Schedule of Recent Experiences: A reliability study. *Psychosomatic Medicine*, 34, 527-531.

Michell, J. (1988). Some problems in testing the double cancellation axiom in conjoint measurement. *Journal of Mathematical Psychology*, 32, 466-473.

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, NJ: Lawrence Earlbaum Associates.

Narens, L. (1985). *Abstract measurement theory.* Cambridge: MIT.

Narens, L., Luce, R.D. (1993). Further comments on the "nonrevolution" arising from axiomatic measurement theory. *Psychological Science*, 4, 127-130.

Pennebaker, J. (1980). *The psychology of physical symptoms.* New York: Springer.

Perline, R., Wright, B.D., and Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, 3, 237-255.

Rasch, G. (1960/80/93). *Probabilistic models for some intelligence and attainment tests.* Chicago: MESA (second reprint).

Riskey, D.R., and Birnbaum, M.H. (1974). Compensatory effects of moral judgment: Two rights do not make up for a wrong. *Journal of Experimental Psychology*, 103, 171-173.

Ross, C.E., and Mirowsky, J. (1979). A comparison of life-event-weighing schemes: Change, undesirability, and effect-proportional indices. *Journal of Health and Social Behavior*, 20, 166-177.

Steele, G., Henderson, S., and Duncan-Jones, D. (1980). The reliability of reporting adverse experiences. *Psychological Medicine*, 10, 301-306.

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16, 433-451.

Thurstone, L.L. (1926). The scoring of individual performance. *Journal of Educational Psychology*, 17, 446-457.

Thurstone, L.L. (1927). The method of paired-comparisons for social values. *Journal of Mathematical and Social Psychology*, 21, 384-400.

Torgerson, W.S. (1958). *Theory and methods of scaling.* New York: Wiley.

Tversky, A. (1967). A general theory of polynomial conjoint measurement. *Journal of Mathematical Psychology*, 4, 1-20.

Wright, B.D., (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, 16, 33-45.

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

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

Appendix A
Notation
Rasch Model Notation

$x_{v\iota}$ response of person v to item ι

For dichotomous responses:

$x_{v\iota} = 0$ Negative response

$x_{v\iota} = 1$ Positive Response (i.e., item endorsement)

For rating scale responses:

$x_{v\iota} = 0, 1, \ldots \kappa$

β_v measure parameter of person v

δ_ι measure parameter of item ι

T_κ step measure parameter of rating scale category κ from $\kappa - 1$

δ_m measure parameter of stimulus m

δ_n measure parameter of stimulus n

δ_{MAX} measure parameter of the maximum item endorsed by person v

e natural log base, e = 2.7182...

An example of a Rasch model probability expression:

$\text{Prob}[x_{v\iota} = 1 \mid \beta_v, \delta_\iota]$ The probability of a positive response, given the person measure (β_v) and the item measure (δ_ι).

Conjoint Measurement Notation

I Row variable

J Column variable

P Dependent variable

a, b, c ∈ I rows a, b, and c are elements of variable I

x, y, z ∈ J columns x, y, z are elements of variable J

Hence, additive conjoint measurement is defined as P being a noninteractive function of I and J, which may be expressed as:

$$P = f(I + J) \text{ or } P = f(I * J).$$

Conjoint measurement in the Rasch framework:

I Person measure variable (β), containing person measure groups as elements

J Item measure variable (ε, the inverse of δ), containing item measure groups as elements

P Proportion of positive responses (in the case of dichotomous responses) obtained by a person measure group to an item measure group

Birnbaum Notation

s_i, s_j, s_k Scale values of stimuli i, j, k, respectively

s_{max} Scale value of the stimulus with the highest scale value

s_{min} Scale value of the stimulus with the lowest scale value

r_{ij} Ratio judgment between stimuli j and i

d_{ij} Difference judgment between stimuli j and i

C_{ijk} Combination judgment of stimuli i, j, and k

w, b_c Constants used to estimate combination judgments

ω Configural weight used to estimate combination judgments

JOURNAL OF OUTCOME MEASUREMENT®, 2(3), 222 - 239

A Research Program for Accountable and Patient-Centered Health Outcome Measures

William P. Fisher, Jr.
Biometry and Genetics
Louisiana State University Medical Center

This article addresses the relevance of probabilistic conjoint (Rasch) measurement to five issues of accountability and patient-centeredness in health care. Goals for research, data quality standards, and standard metrics are proposed. The article is intended to begin to address concerns voiced by health care researchers, policy analysts, and the public about ways in which health care outcome measures can be improved.

Requests for reprints should be sent to William P. Fisher, Jr., Biometry & Genetics, Louisiana State University Medical Center, 1901 Perdido Street, New Orleans, LA 70112

Efforts intended to foster competition in health care on the basis of quality of care have been hampered by issues of data quality. The Health Care Financing Administration (HCFA) publicly disclosed provider-specific mortality rates among hospitalized Medicare patients, but the reports were discontinued because the data were an insufficient basis for making quality inferences (Krakauer, Bailey, Skellan, Stewart, Hartz, Kuhn, & Rimm, 1992; Green, Wintfeld, Sharkey, & Passman, 1990; Green, Passman, & Wintfeld, 1991; "Rating of hospitals is delayed in an effort for stronger data", 1993). The New York State Department of Health's Cardiac Surgery Reporting System (CSRS) was designed to overcome the flaws of the HCFA system. The CSRS was widely regarded as a model system, but its validity has been questioned by clinicians concerned about the accuracy and internal consistency of its data (Topol & Califf, 1994; Green, 1992; Green & Wintfeld, 1994; Green & Wintfeld, 1995; Zinman, 1992).

HCFA, insurance companies, and health management organizations are already tracking provider-specific data, such as mortality rates, and the temptation to make inferences about quality of care from readily-available data, instead of obtaining more relevant but less available data, is nearly irresistible. Many different kinds of measurement systems are in fact by-products of administrative systems. The federal government's family assistance, social welfare, and the Medicare and Medicaid programs often employ administrative forms for the gathering of data in what is in fact a measurement effort. Accordingly, many of the features characteristic of measurement systems, such as 1) clear, standardized definitions, and 2) repeatability within, and reproducibility among, providers, agencies, and regions, are omitted or insufficiently implemented (Bailar, 1985, p. 138; National Research Council, 1983; Forrest, Brown, Scott, Ewy & Flood, 1977).

The situation is no better in the private sector's efforts to devise health outcomes report-cards. A study conducted by the General Accounting Office (GAO) concludes that "no evaluative studies have been conducted to determine the report cards' validity or reliability" (Green & Wintfeld, 1995, p. 1232; *Health care reform: 'report cards' are useful but significant issues need to be addressed*, 1994).

Many of these quality-of-care measuring efforts are based on counts of events, such as mortality, rehospitalization, or the percentage of patients treated who say they are very satisfied with the care they received. It is difficult to evaluate the quality of these distinct data points. Data input errors that do not involve unacceptable characters are almost impossible

to detect using most commonly-employed methods. The validity of such data are questionable because they do not provide any evidence beyond their face value that each use of a code represents another instance of the same thing.

The data's internal consistency should be of special concern, given the GAO's reports of the rarity of reliability and validity studies of report-card outcome reporting systems. Even so, questions regarding the CSRS data's internal consistency arose only because changes were made to the coding instrument (Green & Wintfeld, 1995, p. 1230). Because the data employed in these systems were not intended to measure quality of care, but are taken as a basis for inferences concerning it, some researchers have been motivated to devise rating scale instruments that require less of a leap of faith that the measurement system is measuring what it is supposed to. Multi-question rating scale instruments can overcome data consistency problems by varying the same theme across individual data points, systematically creating the evidence needed for ascertaining that any response to any question is probably based on a consistent interpretation of that question's meaning.

Rating scale measures of health status, functional independence, quality of life, or patient satisfaction could be basic components of accountable, patient-centered health care. Information on patients' perceptions helps clinicians involve patients in their care and improves the quality of care, both of which contribute to enhanced outcomes. Unfortunately, validity and reliability issues associated with unexamined data quality also prevail in rating scale measures (Fisher, 1993; Merbitz, Morris & Grip, 1989; Michell, 1990; Riddick, 1989; Stucki, et al. 1996; Wright & Linacre, 1989; Zhu, 1996). Enhanced quality and precision in health-related measures could deepen and broaden patients' involvement in their care, and could also improve the quality of care.

Probabilistic conjoint (Rasch) measurement (PCM) (Rasch, 1960; Rasch, 1961; Wright, 1968, 1977, 1985; Andrich, 1988; Fisher & Wright, 1994; Smith, 1997) offers some valuable ways of addressing health-related measurement data quality and precision. PCM models test the internal consistency of test or rating scale data, requiring that the measuring function of the instrument remain undisturbed by the particulars of the measurement process, such as the persons or objects measured, the person administering the instrument, and where and when the measure takes place. There is a long history of successful application of these and similar models in education and psychology, dating back to Thurstone's work in the

1920s (Thurstone, 1959). Interest in PCM has grown in recent years with its application in computer-adaptive testing (Lunz, Bergstrom & Gershon, 1994; Reckase, 1989), in multi-rater performance assessment (Linacre, 1989; Linacre, 1996; Linacre, Englehard, Tatum, & Myford, 1994; Englehard, 1992; Tatum, 1991; Myford, 1989), and in quality of life, health status, and functional assessment (Cella, Lloyd, & Wright, 1996; Fisher, A., 1993, 1994; Fisher, A., Bryze, Granger, Haley, Hamilton, et al., 1994; Fisher & Fisher, 1993; Gonin, Lloyd, & Cella, 1996; Haley & Ludlow, 1992; Haley, Ludlow, & Coster, 1993; Haley, McHorney, & Ware, 1994; Harvey & Fisher, 1996; Harvey, Silverstein, Venzon, Kilgore, Fisher, et al., 1992; Heinemann, Linacre, Wright, Hamilton, & Granger, 1993, 1994; Kilgore, Fisher, Silverstein, Harley, & Harvey, 1993; Linacre, Heinemann, Wright, Granger, & Hamilton, 1994; Ludlow, Haley, & Gans, 1992; McHorney, Haley, & Ware, 1997; Silverstein, Kilgore, & Fisher, 1989; Silverstein, Fisher, Kilgore, Harvey, & Harley, 1992; Stucki, Daltroy, Katz, Johannesson, & Liang, 1996; Zhu & Cole, 1996; Zhu & Kurz, 1994).

PCM is valued for its rigorous but flexible data consistency requirements. The flexibility comes from the fact that many kinds of data can be integrated into a single measurement system, but perhaps more important is the probabilistic structures' toleration of random noise. Since rigid, deterministic hierarchical structures, such as Guttman scalograms (Guttman, 1950) set unrealistic expectations for human performance data (Andrich, 1985; Kempen, Myers & Powell, 1995; Wilson, 1989), some other way of modeling behavior, health, and attitudes is needed. Fuzzy logic might seem useful in developing such models, but Rasch's PCM models offer better-understood and more precise properties (Crowther, Batchelder, & Hu, 1995; Fisher, 1995).

Quantitative amounts, by definition, do not depend on the particular sample of persons or items producing them. Tests of sample- and scale-dependency would seem to be crucial to the calibration of quantitative measuring instruments, though these tests are rarely performed (Michell, 1990, 1997). These tests are lacking from the Mokken stochastic scaling employed by Kempen and colleagues (Kempen, et al., 1995), as well as in the more commonly employed method of summated ratings, as has been repeatedly pointed out (Fisher, 1993; Stucki, Daltroy, Katz, et al., 1996; Wright & Linacre, 1989; Zhu, 1996). PCM provides the needed rigor by demanding the consistent hierarchical ordering of items over persons, and vice versa, required for scale-free and sample-free measurement.

Instead of requiring the ratings themselves to exhibit the conjointly

ordered structure, PCM models require it of the probabilities of each person-item interaction. The resulting flexibility helps make sense of the many kinds of existing data, and makes it possible to address new measurement needs quickly and inexpensively. PCM models' rigorous demands for internally consistent data is crucial to the development of health care quality comparisons, since public health care quality measures should be supported by an especially high level of scientific evidence (Epstein, 1995, p. 60).

Instruments capable of consistently producing quantitative effects are crucial to theory development and the gathering of precise, relevant data (Ackermann, 1985; Bud & Cozzens, 1992; Heelan, 1983; Ihde, 1991; van Helden & Hankins, 1994). Tests of the quantitative hypothesis implemented by conjoint models in general (Luce & Tukey, 1964; Michell, 1990), and by PCM models (Brink, 1972; Brogden, 1979; Fisher & Wright, 1994; Perline, Wright, & Wainer, 1979) in particular, determine to what extent instruments measure quantitatively and can therefore focus and support theory development and data gathering.

Five Issues of Accountability and Patient-Centered Care

There are at least five ways of improving health-related outcome measures' accountability and contribution to patient-centered care using PCM. The first three are closely linked: converting ordinal ratings into interval measures via a log-odds transformation (the first improvement) does not make any sense unless the resulting measures are scale free (the second improvement), which in turn provides enhanced meaningfulness and interpretability (the third improvement). The fourth area for improved accountability and patient-centeredness involves computerized administration of self-report rating scale instruments for measuring health status, satisfaction with care, health state preferences, quality of life, etc. Finally, the fifth area concerns consistent contextual influences on measures, such as illness severity or judge-assigned ratings, which can be included in a multifaceted model for estimation and removal from the measures.

Ordinal ratings versus interval measures

The first way of improving health-related outcome measures' accountability and contribution to patient-centered care using PCM begins with the recognition that rating scale data are ordinal. Being ordinal, the unit difference between adjacent scores is probably unknown (Merbitz, et al., 1989; Wright & Linacre, 1989; Michell, 1990), and if it is known, it is

probably highly variable (Stucki, Daltroy, Katz, et al., 1996; Zhu, 1996). All that is known is that a higher score ought to mean that the person measured exhibits more of what is measured than a person with a lower score. Most rating scale measurement methods do not test the hypothesis that the variable of interest is quantitative (Michell, 1990). Interval measures can be made from suitable ordinal data via the log-odds transformation employed in PCM when the internal consistency of the data do not falsify the quantitative hypothesis (Ludlow & Haley, 1995; Wright & Masters, 1982). The log-odds transformation is rarely used in research or management studies employing health-related measures, though its use is increasing.

Scale-dependent versus scale-free measures

Current methods of survey design allow every instrument to measure in its own idiosyncratic unit, making it difficult, if not impossible, to compare measures across instruments, and thus across facilities employing different outcomes measurement systems. The method of summated ratings that predominates in survey design allows the quantitative unit of measurement to vary from instrument to instrument, depending on the number of items and rating scale points involved. Because the meaning of summed scores depends on the particular collection of items administered, these measures are scale-dependent.

Because PCM models can accommodate missing data, the quantitative unit of measurement does not depend on the particular collection of precalibrated items administered. Rasch's PCM models make it possible to equate different instruments so that they measure in the same unit and so that their qualities can be evaluated in the same terms. For instance, the motor skills scales of the Functional Independence Measure (FIM) (Heinemann, Linacre, Wright, et al., 1993; Heinemann, Linacre, Wright, et al., 1994; Linacre, Heinemann, Wright, et al., 1994; Wright, Linacre, & Heinemann, 1993) and the Patient Evaluation Conference System (Harvey, Silverstein, Venzon, et al., 1992; Harvey & Fisher, 1996; Kilgore, Fisher, Silverstein, et al., 1993; Silverstein, Kilgore & Fisher, 1989; Silverstein, Fisher, Kilgore, et al., 1992), two functional assessment instruments widely used in physical medicine and rehabilitation, measure the same construct in the same metric unit, even though the number and content of the instruments' items, and their rating scales, differ (Fisher, Harvey, Taylor, et al., 1995). Similar work has been performed on quality of life measures (Cella, Lloyd & Wright, 1996; Gonin, Lloyd & Cella, 1996), and on patients'

self-reported health status measures (Fisher, Eubanks, & Marier, 1997). Comparisons of four different functional assessment instruments Rasch analyzed on 11 different samples show that these instruments could be co-calibrated to measure in a single metric (Fisher, 1997a, 1997b).

The search for the elusive "gold standard" of outcomes measurement is pointless in the summated ratings environment. The only way to make summated ratings work as a method for standardizing units of measurement is to have each variable universally measured by one collection of items, ensuring that responses are obtained from every person on every item. But even if these improbable steps were taken, there would still be no nonarbitrary metric for the measurement of these variables, since the data will still be ordinal, nonlinear, scale-dependent, and lacking systematic quality control.

In contrast, Rasch's PCM models calibrate scales intended to measure the same variable into nonarbitrary, linear metrics that can be easily equated into a common quantitative system. Furthermore, these models present data quality criteria useful for 1) determining whether different surveys and different survey questions measure the intended variable, 2) comparing instrument reliabilities and validities in a single evaluative framework, and 3) theorizing about the meaning of the variable, which is woefully lacking in most health outcomes measurement, and is also particularly useful in revealing coding or data entry errors.

Enhanced Meaning

Where most approaches to outcome measurement focus on the raw score, a single point of indeterminate meaning, PCM models make it possible 1) to articulate more fully and richly the construct's breadth and depth, and 2) to provide detailed descriptive interpretations of measures (Fisher, Harvey, & Kilgore, 1995; Ludlow & Haley, 1995; Masters, Adams, & Lokan, 1994; Stahl & Lunz, 1996; Woodcock, 1998; Wright, 1977; Wright & Masters, 1982; Zhu & Cole, 1996). Because of PCM models' additive structures, respondent measures and item calibrations are expressed in a common metric that supports interpreting the item difficulty order as delineating a generalized construct relevant to any particular member of the respondent population. The measures can therefore be interpreted in terms of the response probabilities for any calibrated item belonging to the population of items sampled, whether or not that item was actually administered to that respondent.

The importance of enhanced meaning in PCM is such that the phrase

"interpretive measurement" might seem to characterize its major relevant features better than the more often used phrase "objective measurement", which stresses the new form of objectivity achieved in parameter separation. For those researchers who place great importance on the extensive roles played by interpretive strategies in the history of science, and who are furthermore unfamiliar with the extent to which analogues of parameter separation are essential to those interpretive strategies (Fisher, 1992, 1994), the claim to objectivity in measurement is, at best, a quaint anachronism, and at worst, a dire threat to civilization and culture. These researchers might find an interpretively and qualitatively rich quantitative method very attractive, but are put off by the seemingly politically incorrect claim to objectivity and so do not investigate PCM models as thoroughly as they might if the models were more often referred to as facilitating interpretive measurement.

Computer-Enhanced Accuracy, Precision, and Relevance

Current survey methods often stress asking the fewest possible number of questions. The efficiency gained by asking fewer questions is offset by the decreased precision of the information obtained. The effect is to perform a disservice to the patient; people adapt to the needs of the measurement technique, when the measurement technique ought to adapt to the needs of people. Except for one or two isolated instances (Fisher, A., 1993, 1994), computerized administration of rating scale instruments in health care has yet to take advantage of PCM, which has a long history of application to the problems of item banking and adaptive measurement (Choppin, 1968; Wright & Bell, 1984; Lunz, Bergstrom, & Gershon, 1994).

Rasch's probabilistic approach to rating scale measurement makes it possible for surveys, especially computer-automated ones, to include practically the entire universe of questions relevant to a particular variable, since no one respondent need ever be required to answer more questions than they would answer using traditional methods. The respondent's measure is updated with the new information provided by each response, and the next question posed is one with a calibration value within an error of that measure. In this way, more relevant, efficient, and precise measures are made than with traditional approaches.

Severity and Other Adjustments

Other sources of uncontrolled variation introduced into ordinal data's complex of problems prevent use of the data for quality comparisons. These

include illness severity, and rating disagreements among raters. Multifaceted PCM models parameterize and evaluate these sources of variation, making it possible to adjust patient measures and remove the unwanted effects. Applications of multifaceted models (Linacre, 1989; Linacre, Englehard, Tatum & Myford, 1994; Fisher, A., 1993; Fisher, A., 1994; Fisher & Fisher, 1993; Lunz & Stahl, 1993a, 1993b; Myford, 1989; Tatum, 1991; Englehard, 1992) to performance assessment data show that variation in measures can more often be due to the particular rater involved than to performance-based differences among those measured. When ratings assigned by judges observing a behavior or a performance are statistically consistent with each other, it no longer matters whether the individual judges' ratings disagree, since it is then a simple matter to remove judge-specific deficits or gains from the measures by adding or subtracting the judges' calibrations from the measures.

It may similarly be possible to effect illness severity adjustments through multifaceted PCM, though no or very little research in this area has yet been done.

Goals for Research and Standards

The increasing need for accountability and efficiency in health care makes these oversights in accountability for outcomes measures intolerable. As competition among providers intensifies, those best able to identify and act on quality and efficiency issues will have the advantage; the superiority of measures based on PCM models may be decisive in some markets. As medical records become computer-based, and as computer-based data expands beyond individual machines to networks within facilities and then to networks of facilities, health status and patient satisfaction measurement systems that can critically evaluate and flexibly incorporate various question phrasings and rating scale formats into a single metric can only increase in demand.

Health care reform cannot realize its goals as long as its accountability is based on data that are themselves not accountable. Quality of care cannot be assessed and improved when the numbers treated as measures literally do not add up; when measures are not interpreted as ranges bound by error but as errorless points; when measures are not comparable across instrument brands, samples of patients, and the facilities using them; and when the choice of a measure is determined more by the dictates of an archaic and obsolete methodology than by the needs of patients and their health care providers. None of these obstacles to improved accountability

and patient-centeredness involves difficult conceptual dilemmas or technical demands requiring extensive research. The problems we face are much more matters of organization and education.

Recent work in social studies of science stresses that the objectivity of physical measurement follows less from some mysterious and special "natural" quality associated with the things measured than it does from the existence of professional organizations of metrologists dedicated to establishing, monitoring, and disseminating common units of measurement (Bud & Cozzens, 1992; O'Connell, 1993). Health outcomes researchers are faced with a choice as to which metrological system they want to use for establishing a common currency for the exchange of health outcomes values. One choice is based on the nonlinear, incomparable, and inefficient raw scores derived from summated ratings, which require everyone who wants to participate in the measurement system to use the same instrument and to obtain complete data from every respondent. Another choice is based on the linear, comparable, and efficient measures derived from application of PCM models, models that support interval-level measurement; provide error and statistical consistency estimates for every respondent and item; make it possible to equate different instruments to measure in a single metric unit; enhance the meaningfulness and interpretability of measures; adapt instruments to the needs of people, instead of vice versa; and facilitate estimation and removal of consistent, but unwanted, effects on measures, such as rater harshness/leniency, or illness severity.

Explicit standards for measurement quality, similar to those spelled out by Hunter (1980) for physical measurement, are being articulated (Fisher, 1997b). In brief, these standards specify several statistical criteria that instruments measuring particular variables have to meet or surpass in order to be certified as measuring in the relevant unit. These statistics for instrument certification will probably include, but will not be limited to 1) a correlation of at least .85 between the tested instrument's measures and the reference instrument's measures of the construct on a common sample; 2) a correlation of at least .85 between any items on the tested instrument and any on the reference instrument identified as addressing the same aspect of the construct; 3) a reliability coefficient of at least .85 (which indicates a ratio of variation to error of about 2.6 to 1); and 4) sufficient indication of statistical consistency (data-model fit) on the part of both the item calibration estimates and the person measures. Similar statistics on the instrument's performance on diverse samples taken from geographically separated locations and different kinds of facilities also need to be

considered.

New instruments set reference standards when they establish the best values for the measurement range, variation, error, ratio of variation to error, or data-model fit. It may happen that no one brand of instrument becomes the reference standard for a variable. Perhaps it would be useful to consider the entire collection of items from certified instruments as the reference standard.

Next Steps

These and related questions are debated in the International Outcome Measurement Conferences (Smith, 1997), most recently held at the University of Chicago in May, 1998. In addition to presenting recent advances in the application of Rasch measurement models to health outcomes variables, this group is beginning to address issues concerning standards criteria for specific variables; the practical logistics of how instruments will be tested and certified; which instruments will be certified as initial reference standards; the creation of an independent metrology group for monitoring and enforcing standards; and the publication of a peer-reviewed journal for documenting these activities and research associated with them. All interested parties are invited to participate.

Acknowledgments

The author would like to acknowledge the influence and support of Robert Marier, James Diaz, Miguel Guzman, Theo Dawson, and the late Gordon Black. Thanks also to Richard Scribner for his comments on an earlier version of this paper. As always, Benjamin Wright's impact on this work remains incalculable.

References

Ackermann, J. R. (1985). *Data, instruments, and theory: A dialectical approach to understanding science.* Princeton, New Jersey: Princeton University Press.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. B. Tuma (Ed.), *Sociological methodology 1985* (pp. 33-80). San Francisco: Jossey-Bass.

Andrich, D. (1988). *Rasch models for measurement.* Sage University Paper Series on Quantitative Applications in the Social Sciences, vol. series no. 07-068. Beverly Hills, California: Sage Publications.

Bailar, B. A. (1985). Quality issues in measurement. *International Statistical Review, 53*(2), 123-139.

Brink, N. E. (1972). Rasch's logistic model vs the Guttman model. *Educational and Psychological Measurement, 32,* 921-927.

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika, 42,* 631-634.

Bud, R., & Cozzens, S. E. (Editors). (1992). *SPIE Institutes. Vol. 9: Invisible connections: instruments, institutions, and science* (R. F. Potter, Ed.). Bellingham, WA: SPIE Optical Engineering Press.

Cella, D. F., Lloyd, S. R., & Wright, B. D. (1996). Cross-cultural instrument equating: Current research and future directions. In B. Spilker (Ed.), *Quality of life and pharmacoeconomics in clinical trials (2d edition)* (pp. 707-715). New York, New York: Lippincott-Raven.

Choppin, B. (1968). An item bank using sample-free calibration. *Nature, 219,* 870-872.

Crowther, C. S., Batchelder, W. H., & Hu, X. (1995, April). A measurement-theoretic analysis of the fuzzy logic model of perception. *Psychological Review, 102*(2), 396-408.

Englehard, G., Jr. (1992). The measurement of writing ability with a many-facet Rasch model. *Applied Measurement in Education, 5*(3), 171-191.

Epstein, A. (1995). Performance reports on quality: Prototypes, problems, and prospects. *New England Journal of Medicine, 333*(1), 57-61.

Fisher, A. G. (1993, April). The assessment of IADL motor skills: An application of many-faceted Rasch analysis. *American Journal of Occupational Therapy, 47*(4), 319-329.

Fisher, A. G. (1994). Development of a functional assessment that adjusts ability measures for task simplicity and rater leniency. In M. Wilson (Ed.), *Objective measurement: Theory into practice. Vol II* (pp. 145-175). Norwood, New Jersey: Ablex Publishing Corporation.

Fisher, A. G., Bryze, K. A., Granger, C. V., Haley, S. M., Hamilton, B. B., Heinemann, A. W., Puderbaugh, J. K., Linacre, J. M., Ludlow, L. H., McCabe, M. A., & Wright, B. D. (1994). Applications of conjoint measurement to the development of functional assessments. *International Journal of Educational Research, 21*(6), 579-593.

Fisher, W. (1992). Objectivity in measurement: A philosophical history of Rasch's separability theorem. In M. Wilson (Ed.), *Objective Measurement: Theory into practice*, Vol. I. Norwood, NJ: Ablex Publishing Corp. (pp. 29-58).

Fisher, W. P., Jr. (1993). Measurement-related problems in functional assessment. *The American Journal of Occupational Therapy, 47*(4), 331-338.

Fisher, W. P., Jr. (1994). The Rasch debate: Validity and revolution in educational measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice*. Vol. II. Norwood, NJ: Ablex Publishing Corp. (pp. 36-72).

Fisher, W. P., Jr. (1995). Fuzzy truth and the Rasch model. *Rasch Measurement Transactions, 9*(3), 442.

Fisher, W. P., Jr. (1997a). Physical disability construct convergence across instruments: Towards a universal metric. *Journal of Outcome Measurement, 1*(2), 87-113.

Fisher, W. P., Jr. (197b, June). What scale-free measurement means to health outcomes research. *Physical Medicine & Rehabilitation State of the Art Reviews, 11*(2), 357-373.

Fisher, W. P., Jr., Eubanks, R. L., & Marier, R. L. (1997). Equating the MOS SF36 and the LSU HSI physical functioning scales. *Journal of Outcome Measurement, 1*(4), 329-362.

Fisher, W. P., Jr., & Fisher, A. G. (1993). Applications of Rasch analysis to studies in occupational therapy. C. V. Granger & G. E. Gresham (Eds.), *New developments in functional assessment: Physical Medicine and Rehabilitation Clinics of North America, 4*(3), 551-569.

Fisher, W. P., Jr., Harvey, R. F., & Kilgore, K. M. (1995). New developments in functional assessment: Probabilistic models for gold standards. *NeuroRehabilitation, 5*(1), 3-25.

Fisher, W. P., Jr., Harvey, R. F., Taylor, P., Kilgore, K. M., & Kelly, C. K. (1995). Rehabits: A common language of functional assessment. *Archives of Physical Medicine and Rehabilitation, 76,* 113-122.

Fisher, W. P., Jr., & Wright, B. D. (1994). Introduction to probabilistic conjoint measurement theory and applications. *International Journal of Educational Research, 21*(6), 559-568.

Forrest, W. H., Jr., Brown, B. W., Jr., Scott, W. R., Ewy, W., & Flood, A. B. (1977). *Determinants of service intensity in the medical care sector.* Washington, DC: National Academy of Sciences.

Gonin, R., Lloyd, S. R., & Cella, D. F. (1996). Establishing equivalence between scaled measures of quality of life. *Quality of Life Research,* pp. 20-26.

Green, J. (1992, Spring). Problems in the use of outcome statistics to compare health care providers. *Brooklyn Law Review, 58,* 55-73.

Green, J., Passman, L., & Wintfeld, N.. (1991). Analyzing hospital mortality: The consequences of diversity in patient mix. *Journal of the American Medical Association, 265,* 1849-1853.

Green, J., & Wintfeld, N. (1994). *Report cards on cardiac surgeons: Assessing New York State's approach: Final report to the United Hospital Fund.* New York: New York University Medical Center.

Green, J., & Wintfeld, N. (1995). Report cards on cardiac surgeons: Assessing New York State's approach. *New England Journal of Medicine, 332*(18), 1229-1232.

Green, J., Wintfeld, N., Sharkey, P., & Passman, L. (1990). The importance of severity of illness in assessing hospital mortality. *Journal of the American Medical Association, 263,* 241-246.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer & et al. (Eds.), *Studies in social psychology in World War II. volume 4: Measurement and prediction* (pp. 60-90). New York: Wiley.

Haley, S. M., & Ludlow, L. H. (1992). Applicability of the hierarchical scales of the Tufts Assessment of Motor Performance for school-aged children and adults with disabilities. *Physical Therapy, 72*(3), 191-202.

Haley, S. M., Ludlow, L. H., & Coster, W. J. (1993). Pediatric Evaluation of Disability Inventory: Clinical interpretation of summary scores using Rasch rating scale methodology. C. V. Granger & G. E. Gresham (Eds.), *New developments in functional assessment: Physical Medicine and Rehabilitation Clinics of North America, 4*(3), 529-540

Haley, S. M., McHorney, C. A., & Ware, J. E., Jr. (1994). Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. unidimensionality and reproducibility of the Rasch item scale. *Journal of Clinical Epidemiology, 47*(6), 671-684.

Harvey, R. F., & Fisher, W. P., Jr. (1996). The Patient Evaluation Conference System (PECS©). In J. McGee, N. Goldfield, J. Morton & K. Riley (Eds.), *Collecting Information from Patients: A Resource Manual of Tested Questionnaires and Practical Advice.* Gaithersburg, Maryland: Aspen Publications, Inc.

Harvey, R. F., Silverstein, B., Venzon, M. A., Kilgore, K. M., Fisher, W. P., Jr., Steiner, M., & Harley, J. P. (1992). Applying psychometric criteria to functional assessment in medical rehabilitation: III. construct validity and predicting level of care. *Archives of Physical Medicine and Rehabilitation, 73*(10), 887-892.

Health care reform: 'report cards' are useful but significant issues need to be addressed [GAO/HEHS-94-219]. (1994). Washington, DC: General Accounting Office.

Heelan, P. (1983, June). Natural science as a hermeneutic of instrumentation. *Philosophy of Science, 50,* 181-204.

Heinemann, A. W., Linacre, J. M., Wright, B. D., Hamilton, B. B., & Granger, C. V. (1993). Relationships between impairment and physical disability as measured by the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation, 74*(6), 566-573.

Heinemann, A. W., Linacre, J. M., Wright, B. D., Hamilton, B. B., & Granger, C. V. (1994). Prediction of rehabilitation outcomes with disability measures. *Archives of Physical Medicine and Rehabilitation, 75*(2), 133-143.

Hunter, J. S. (1980, November). The national system of scientific measurement. *Science, 210*(21), 869-874.

Ihde, D. (1991). *Instrumental realism: The interface between philosophy of science and philosophy of technology.* The Indiana Series in the Philosophy of Technology. Bloomington, Indiana: Indiana University Press.

Kempen, G. I., J M, Myers, A. M., & Powell, L. E. (1995). Hierarchical structure in ADL and IADL: Analytical assumptions and applications for clinicians and researchers. *Journal of Clinical Epidemiology, 48*(11), 1299-1305.

Kilgore, K. M., Fisher, W. P., Jr., Silverstein, B., Harley, J. P., & Harvey, R. F. (1993). Application of Rasch analysis to the Patient Evaluation and Conference System. C. V. Granger & G. E. Gresham (Eds.), *New developments in functional assessment. Physical Medicine and Rehabilitation Clinics of North America, 4*(3), 493-515.

Krakauer, H., Bailey, R. C., Skellan, K. J., Stewart, J. D., Hartz, A. J., Kuhn, E. M., & Rimm, A. A. (1992). Evaluation of the HCFA model for the analysis of mortality following hospitalization. *Health Services Research, 27*(3), 317-335.

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

Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. In G. Englehard, Jr. & M. Wilson (Eds.), *Objective measurement: Theory into practice, vol. 3* (pp. 85-98). Norwood, NJ: Ablex Publishing Co.

Linacre, J. M., Englehard, G., Tatum, D. S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research, 21*(6), 569-577.

Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V., & Hamilton, B. B. (1994). The structure and stability of the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation, 75*(2), 127-132.

Ludlow, L. H., & Haley, S. M. (1995, December). Rasch model logits: Interpretation, use, and transformation. *Educational and Psychological Measurement, 55*(6), 967-975.

Ludlow, L. H., Haley, S. M., & Gans, B. M. (1992). A hierarchical model of functional performance in rehabilitation medicine: The Tufts Assessment of Motor Performance. *Evaluation & the Health Professions, 15,* 59-74.

Lunz, M. E., Bergstrom, B. A., & Gershon, R. C. (1994). Computer adaptive testing. *International Journal of Educational Research, 21*(6), 623-634.

Lunz, M. E., & Stahl, J. A. (1993a). Impact of examiners on candidate scores: An introduction to the use of multifacet Rasch model analysis for oral examinations. *Teaching and Learning in Medicine, 5*(3), 174-181.

Lunz, M. E., & Stahl, J. A. (1993b, April). The effect of rater severity on person ability measures: A Rasch model analysis. *American Journal of Occupational Therapy, 47*(4), 311-317.

McHorney, C. A., Haley, S. M., & Ware, J. E. (1997). Evaluation of the MOS SF-

36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. *Journal of Clinical Epidemiology, 50*(4), 451-461.

Masters, G. N., Adams, R., & Lokan, J. (1994). Mapping student achievement. *International Journal of Educational Research, 21*(6), 595-609.

Merbitz, C., Morris, J., & Grip, J. (1989). Ordinal scales and the foundations of misinference. *Archives of Physical Medicine and Rehabilitation, 70,* 308-312.

Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*(3), 355-383.

Myford, C. M. (1989). The nature of expertise in aesthetic judgment: Beyond inter-judge agreement [Diss]. *Dissertation Abstracts International, 50,* 3562A, Chicago, Illinois: University of Chicago.

National Research Council. (1983). *Family assistance and poverty: An assessment of statistical needs.* Washington, DC: National Academy of Sciences.

O'Connell, J. (1993). Metrology: The creation of universality by the circulation of particulars. *Social Studies of Science, 23,* 129-173.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*(2), 237-255.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (pp. 321-333). Berkeley, California: University of California Press.

Rating of hospitals is delayed in an effort for stronger data. (1993, June 23). *New York Times,* p. A19.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8,* 3.

Riddick, L. (1989). Quantitative assessments: Their mathematical faults. *WFOT Bulletin, 20,* 11-13.

Silverstein, B. J., Fisher, W. P., Jr., Kilgore, K. M., Harvey, R. F., & Harley, J. P. (1992). Applying psychometric criteria to functional assessment in medical rehabilitation: II. defining interval measures. *Archives of Physical Medicine and Rehabilitation, 73*(6), 507-518.

Silverstein, B. J., Kilgore, K. M., & Fisher, W. P., Jr. (1989). *Implementing patient tracking systems and using functional assessment scales.*Center for Rehabilitation Outcome Analysis monograph series on issues and methods in rehabilitation outcome analysis, Vol. 1. Wheaton, Illinois: Marianjoy Rehabilitation Center.

Smith, R. M. (Editor). (1997, June). *Physical Medicine & Rehabilitation State of the Art Reviews, 11*(2).*Proceedings of the First International Outcome Measurement Conference*. Hanley & Belfus.

Stahl, J. A., & Lunz, M. E. (1996). Judge performance reports: Media and message. In G. Englehard & M. Wilson (Eds.), *Objective Measurement: Theory into Practice, Volume 3* (pp. 113-125). Norwood, NJ: Ablex.

Stucki, G., Daltroy, L., Katz, N., Johannesson, M., & Liang, M. H. (1996). Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. *Journal of Clinical Epidemiology, 49*(7), 711-717.

Tatum, D. S. (1991). A measurement system for speech evaluation [Diss]. *Dissertation Abstracts International, 52*(4), 1301A, Chicago, Illinois: University of Chicago.

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

Topol, E., & Califf, R. (1994). Scorecard cardiovascular medicine: Its impact and future directions. *Annals of Internal Medicine, 120,* 65-70.

Van Helden, A., & Hankins, T. L. (Eds.). (1994). *Instruments. A Special Issue of Osiris: A Research Journal Devoted to the History of Science and Its Cultural Influences, vol. 9.* Chicago: University of Chicago Press.

Wilson, M. (1989). A comparison of deterministic and probabilistic approaches to learning structures. *Australian Journal of Education, 33*(2), 127-140.

Woodcock, R. (1998). What can Rasch-based scores convey about a person's test performance? In S. Embretson & S. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (p. in press). Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 invitational conference on testing problems* (pp. 85-101). Princeton, New Jersey: Educational Testing Service.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97-116.

Wright, B. D. (1985). Additivity in psychological measurement. In E. Roskam (Ed.), *Measurement and personality assessment.* North Holland: Elsevier Science Ltd.

Wright, B. D., & Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement, 21*(4), 331-345.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*(12), 857-867.

Wright, B. D., Linacre, J. M., & Heinemann, A. W. (1993). Measuring functional status in rehabilitation. C. V. Granger & G. E. Gresham (Eds.), *New developments in functional assessment. Physical Medicine and Rehabilitation Clinics of North America, 4*(3), 475-491.

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

Zhu, W. (1996). Should total scores from a rating scale be used directly? *Research Quarterly for Exercise and Sport, 67*(3), 363-372.

Zhu, W., & Cole, E. L. (1996). Many-faceted Rasch calibration of a gross-motor instrument. *Research Quarterly for Exercise and Sport, 67*(1), 24-34.

Zhu, W., & Kurz, K. A. (1994). Rasch partial credit analysis of gross motor competence. *Perceptual and Motor Skills, 79,* 947-961.

Zinman, D. (1992, May 12). Surgery by odds: Risky heart cases rejected. *New York Newsday,* p. 87.

JOURNAL OF OUTCOME MEASUREMENT®, 2(3), 240 - 265

Measuring Individual Differences in Change with Multidimensional Rasch Models

Wen-chung Wang
National Chung-Cheng University

Mark Wilson
University of California, Berkeley

Raymond J. Adams
Australian Council for Educational Research

Item response models have been developed to explore change measurement, including those proposed by Fischer and his colleagues (e.g., Fischer & Pazer, 1991; Fischer & Ponocny, 1994), Andersen (1985) and Embretson (1991). In this article, we propose another multidimensional Rasch model, the multidimensional random coefficient multinomial logit (MRCML) model (Adams, Wilson, & Wang, 1997). All these models are briefly reviewed and compared. The MRCML can be applied to not only polytomous items but also investigation of variations in item difficulties. Based on variations in difficulties across occasions and items, five kinds of models are proposed. Some simulation studies were conducted to examine parameter recovery of the MRCML model under various testing situations. All the parameters were recovered very well. A real data set was analyzed to show applications of the MRCML to measuring individual differences in change.

Requests for reprints should be sent to Wen-chung Wang, Department of Psychology, National Chung-Cheng University, Chia-Yi, Taiwan.

Measuring change has long been an interest of modern psychological measurement. Two major methods are generally used: (a) A gain or a difference score by subtracting the pretest score from the posttest score;(b) The pretest score is treated as a covariate and the posttest score as a criterion variable. Some researchers argued that the covariance approach is superior (e.g., Cohen & Cohen, 1975; Cronbach & Furby, 1970), whereas some preferred the gain score approach (e.g., Gottman & Krokoff, 1990; Wainer, 1991). The gain score approach has an advantage of more intuitive and direct estimates of change. However, as Bereiter (1963) pointed out, there are three problems when change is measured on the basis of gain scores: (a) The reliability of gain scores between two tests is inversely related to the correlation between the two tests; (b) For subjects with different initial scores, their gain scores may not be on the same scale; (c) Gain scores are generally negatively correlated with baseline levels. Even if no real gain occurs, the negative correlation still exists because of the effect of "regression toward the mean" (Lord, 1963). Some researchers have been working on these problems (e.g., Geenen & van de Vijver, 1993; Jamieson, 1993, 1994; Jin, 1992; Llabre, Spitzer, Saab, Ironson, & Schneiderman, 1991; Malgady & Colon-Malgady, 1991), nevertheless, these problems seem irresolvable because the change measurements are based on classical test theory, which is widely known as population specific.

Recent developments in item response theory (IRT) have explored the measurement of change. For example, Fischer and Pazer (1991) have proposed a linear rating scale model with an application to the measurement of change. The model is an extension of the rating scale model (Andrich, 1978). Following this line, Fischer and Ponocny (1994) have extended the partial credit model (Masters, 1982) to a linear partial credit model. They have also shown how the linear partial credit model can be applied to the measurement of change.

Andersen (1985) has also shown a measurement model for longitudinal latent structure between repeated testings. The model combines the values of the latent variable at several time points (or conditions) in a multidimensional latent density and directly estimates the variance-covariance matrix among the values. Likewise, Embretson (1991) has presented a multidimensional latent model for measuring learning and change. She postulated a simplex structure to link item responses to an initial ability and one or more modifiabilities (or learning abilities). The model, unlike Andersen's, decomposes the *effective* ability involved in the latter

occasion into an initial ability and one or more modifiabilities, which represent individual change across occasions.

All the above models have some drawbacks. First, although the linear partial and the linear rating scale model can be applied to polytomous items, they are not suitable for measuring individual differences in change because all individuals are assumed to change in the same amount across occasions. Second, even though individual differences in change can be measured by using both Andersen's and Embretson's models, they are limited to dichotomous items. Third, both Andersen's and Embretson's models assume the item difficulty remains unchanged across occasions, which is a rather strict assumption because items might express variations in difficulties at different occasions due to practice, memory, or response consistency effects. In such cases, we would like to model these complex structures to correspond to real testing situations and to examine variation in item difficulties across occasions. A recently developed multidimensional random coefficients multinomial logit (MRCML) model (Adams, Wilson, & Wang, 1997) can meet the demand.

In this article, we first (a) briefly review Fischer and his colleagues' linear partial credit model, Andersen's, and Embretson's models for measurement of change, (b) introduce the MRCML and show how it relates to the three models, (c) demonstrate how the MRCML can be applied to polytomous items and to investigation of variation in item difficulties across occasions, (d) conduct simulation studies to examine parameter recovery of the MRCML under various testing situations for the measurement of change, and finally (e) analyze a real data set measuring changes of student-family and student-peer relationships before and after subjects left homes for college education.

The Linear Partial Credit Model

Because the linear rating scale model is a special case of the linear partial credit model, just as the rating scale model is a special case of the partial credit model, only the linear partial credit model is described. The linear partial credit model is an extension of the partial credit model, which can be expressed as

$$P\left(X_{nij}=1\middle|\theta_n,\delta_{ij}\right)=\frac{exp\left(j\theta_n+\delta_{ij}\right)}{\sum_{t=0}^{m_i} exp\left(t\theta_n+\delta_{ij}\right)} \qquad (1)$$

for $i = 1, ..., I$ items and $j = 0, ..., m_i$ response categories. X_{nij} is an indicator variable taking a value of 1 if person n falls into response category j of item i and a value of 0, otherwise; δ_{ij} is the easiness of response category j of item i; θ_n is the ability level of person n.

In the linear partial credit model, δ_{ij} is further partitioned into

$$\delta_{ij} = \sum_{l=1}^{p} w_{ijl}\alpha_l + jc, \tag{2}$$

where c is a normalization constant, α_l, for $l = 1, ..., p$, are "basic parameters" measuring the effects of structural item properties or of experimental conditions on the response, and w_{ijl} are predefined weights of the α_l.

The linear partial credit model is more general than the partial credit model, in formal respect.However, with respect to empirical data, the linear partial credit model is more restrictive, because the linear constraints typically reduce the number of free parameters.This is also true for the linear rating scale model to the rating scale model.Fischer and Ponocny (1994) have applied the linear partial credit model to a data set from a clinical study on patients with certain psychosomatic disorders to assess the effects of relaxation training and trend.They have concluded that the linear partial credit model is a useful and practical tool for analyzing polytomous data, particularly so for testing hypotheses on change.

In the linear partial credit model as well as in the linear rating scale model, the trend and treatment effects are assumed to be equal for all persons who receive the same treatment or time intervals.Put another way, the change is assumed to be identical across individuals.Therefore, the linear partial credit model is not modeling individual differences in change.

Andersen's Multidimensional Rasch Model for Repeated Testings

Consider dichotomous items administered to the same examinees across K occasions or conditions (e.g., longitudinal data).The item difficulties are assumed to be constant across occasions, but the abilities involved depend on the specific occasion. According to Andersen's model, the probability of passing item i for person n at occasion k follows the Rasch dichotomous model as

$$P(X_{nik} = 1|\theta_{nk}, b_i) = \frac{exp(\theta_{nk} - b_i)}{1 + exp(\theta_{nk} - b_i)}, \tag{3}$$

where θ_{nk} is the ability of person n at occasion k and b_i is the difficulty of item i.

Equation (3) leads to:

$$log\left(\frac{P\left(X_{nik}=1|\theta_{nk},b_i\right)}{P\left(X_{nik}=0|\theta_{nk},b_i\right)}\right)=\theta_{nk}-b_i, \tag{4}$$

where $p(X_{nik} = 1)$ and $p(X_{nik} = 0)$ (θ_{nk} and b_i are omitted if not confusing) denote the probabilities of being correct and incorrect, respectively, for person n to item i at occasion k. Note that the item difficulty b_i does not depend on k, meaning that the difficulty is unchanged across occasions. Although each item has only one difficulty parameter across occasions, each person has one ability parameter on each occasion.

The correlations among K dimensions can be estimated directly to depict the relationships of the underlying latent variable across occasions. Although there are K abilities, Andersen characterized the latent variable as unidimensional measured at different time points. In this respect, the model is unidimensional. Andersen was more interested in the underlying latent correlation than in measuring individual differences in change. The change can only be implicitly inferred through the relationship among the K abilities and by subtracting θ_{nk} from θ_{nk+1}.

With respect to the measurement of individual differences in change, Andersen's model is one step ahead than the linear partial credit model, in that the former is appropriate for understanding the impact of time or treatment on the ability distribution. However, abilities in Andersen's model are occasion-specific, as the θ_{nk} shown in Equations (3) and (4), because no change parameters for individuals are set up in the model. In addition, Andersen's model was limited to the repeated administration of the same dichotomous items.

Embretson's Multidimensional Rasch Model for Learning and Change

It is possible that the change parameters for individuals are built in measurement models. Embretson (1991) has provided a multidimensional Rasch model for learning and change, which is appropriate for ability measurements repeated either under varying conditions or upon different occasions. Furthermore, it can be applied to situations where items are not repeated. In Embretson's model, the probability of passing item i for person n at occasion k is expressed as

$$P\left(X_{nik}=1\middle|\boldsymbol{\theta}_n,b_i\right)=\frac{exp\left(\sum_{m=1}^{k}\theta_{nm}-b_i\right)}{1+exp\left(\sum_{m=1}^{k}\theta_{nm}-b_i\right)}, \tag{5}$$

where θ_{nm} is the additional ability needed of person n at occasion m, and others are defined as those in Andersen's model.

Equation (5) leads to:

$$log\left(\frac{P\left(X_{nik}=1\middle|\boldsymbol{\theta}_n,b_i\right)}{P\left(X_{nik}=0\middle|\boldsymbol{\theta}_n,b_i\right)}\right)=\sum_{m=1}^{k}\theta_{nm}-b_i. \tag{6}$$

The item difficulty remains constant across occasions, as in Andersen's model. At the first occasion, a person needs an initial ability, θ_{n1}, to pass the item. At the second occasion, in addition to the initial ability an extra effort, the first modifiability, θ_{n2}, is needed to pass the item. At the third occasion, in addition to the initial ability and the first modifiability, another extra effort, the second modifiability, θ_{n3}, is needed to pass the item, and so on. Therefore, the modifiability can be treated as the change (or learning) between two successive occasions.

Although Andersen's and Embretson's models are called "multidimensional" because each person has more than one ability parameter, there is only one single latent trait measured. Actually, this unidimensionality is requirement for any change measurement. Suppose each occasion aims at a distinct dimension, no change can be measured. Put another way, θ_{nk} and θ_{nk+1} in Andersen's model belong to the same latent trait but at different occasions. All the initial ability and the successive modifiabilities in Embretson's model belong to the same trait, too. Therefore, we can conceptualize both models as unidimensional but multi-parameter.

The MRCML

The MRCML is a multidimensional extension of the random coefficients multinomial logit model (Adams and Wilson, 1996). Assume that a set of D traits underlie the individuals' responses and the individuals' positions

are represented by the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_D)'$. Suppose we have I items indexed $i = 1, ..., I$ with each item admitting $K_i + 1$ response alternatives indexed $k = 0, 1, ..., K_i$. We then use the vector valued random variable $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2}, ..., \mathbf{X}_{ikl})'$, where

$$X_{ik} = \begin{cases} 1 & \text{if response to item } i \text{ is in category } k, \\ 0 & \text{otherwise,} \end{cases}$$

to indicate the $K_i + 1$ possible responses to item i.

A response in the zero category is denoted by a vector of zeroes. This effectively makes the zero category a reference category and is necessary for model identification. The choice of this as the reference category is arbitrary and does not affect the generality of the model. For the Rasch model, the incorrect response is usually treated as the reference category.

The items are described through a vector $\boldsymbol{\xi} = (\xi_1, \xi_2, ..., \xi_p)'$, of p parameters.Linear combinations of these are used in the response probability model to describe the empirical characteristics of the response categories of each item. These linear combinations are defined by design vectors $\mathbf{a}_{ik}$, $(i = 1, ..., I; k = 1, ... K_i)$ each of length p which can be collected to form a design matrix $\mathbf{A} = (\mathbf{a}_{11}, \mathbf{a}_{12}, ..., \mathbf{a}_{1K1}, \mathbf{a}_{21}, ..., \mathbf{a}_{2k2}, ..., \mathbf{a}_{1K1})'$.

An additional feature of the MRCML is the introduction of a scoring function that allows the specification of the score or 'performance level' that is assigned to each possible response to each item. A response in category k in dimension d of item i is scored b_{ikd}. The scores across D dimensions can be collected into a column vector $\mathbf{b}_{ik} = (b_{ik1}, b_{ik2}, ..., b_{ikD})'$, (By definition, the score for a response in the zero category is zero, but other responses may also be scored zero.) and again be collected into a scoring sub-matrix for item i, $\mathbf{B}_i = (b_{i1}, b_{i2}, ..., b_{iD})'$, and then collected into a scoring matrix $\mathbf{B} = (B'_1, B'_2, ..., B'_I)'$ for the whole test.

In the MRCML the probability of a response in category k of item i for person n is modeled as

$$P(X_{nik} = 1|\boldsymbol{\theta}_n, \boldsymbol{\xi}) = \frac{exp(\mathbf{b}'_{ik}\boldsymbol{\theta}_n + \mathbf{a}'_{ik}\boldsymbol{\xi})}{\sum_{j=1}^{K_i} exp(\mathbf{b}'_{ij}\boldsymbol{\theta}_n + \mathbf{a}'_{ij}\boldsymbol{\xi})}. \quad (7)$$

Note that the item score vector $\mathbf{b}_{ik}$ in the MRCML is not a free parameter, but is known a priori. Accordingly, the MRCML belongs to the family of Rasch measurement models (Rasch, 1960/1980, 1961), so that interpretation of the item parameters is simpler than for models where discrimination parameters are present.

Comparing the MRCML with the linear partial credit model, as shown in Equations (1) and (2), we can find that the MRCML is a super-model of the linear partial credit model, although somewhat different notations are used. In fact the linear partial credit model is a special case of the unidimensional random coefficients multinomial logit model, which is a unidimensional version of the MRCML. We do not go into details in this article. Interested readers are referred to Adams and Wilson (1996). Also, because the linear partial credit model is not modeling individual differences in change, we focus on how the MRCML comprises Andersen's and Embretson's models as special cases under various hypothetical testing situations.

How the MRCML Integrates Andersen's and Embretson's Models

Comparing Andersen's model (Equation (4)) with Embretson's (Equation (6)), we find that

$$\theta_{Ak} = \sum_{m=1}^{k} \theta_{Em},$$

where subscript A and E denote Andersen's and Embretson's model, respectively. More over,

$$\theta_{E1} = \theta_{A1},$$

$$\theta_{E2} = \theta_{A2} - \theta_{A1},$$

$$\theta_{Ek} = \theta_{Ak} - \theta_{Ak-1}.$$

Accordingly, their means have the following relationship.

$$\bar{\theta}_{E1} = \bar{\theta}_{A1},$$

$$\bar{\theta}_{E2} = \bar{\theta}_{A2} - \bar{\theta}_{A1},$$

$$\bar{\theta}_{Ek} = \bar{\theta}_{Ak} - \bar{\theta}_{Ak-1}. \tag{8}$$

Regarding the variance-covariance matrix of the person ability distribution for the two models, it can be easily shown that:

$$\begin{aligned}
&\sigma^2(\theta_{E1}) = \sigma^2(\theta_{A1}),\\
&\sigma^2(\theta_{E2}) = \sigma^2(\theta_{A2}) + \sigma^2(\theta_{A1}) - 2\sigma(\theta_{A2}, \theta_{A1}),\\
&\sigma(\theta_{E2}, \theta_{E1}) = \sigma(\theta_{A2}, \theta_{A1}) - \sigma^2(\theta_{A1}).
\end{aligned} \tag{9}$$

The ideas of these two models are quite similar. The change in Andersen's model is investigated through the covariances among abilities across occasions. This approach is analogous to that of correlating posttest scores to pretest scores. In contract, Embretson subtracted the preceding ability from its subsequent ability to form the gain, which she called the modifiability. This approach is analogous to that of subtracting pretest scores from posttest scores to form gain scores.

We next show how the MRCML incorporates these two models by manipulating the scoring and the design matrices. It should be noted that the designing matrices for Andersen's and Embretson's are in fact identical. It is the scoring matrix that makes the two models different. For simplicity, let two dichotomous items be administered across three occasions. Figure 1 shows the scoring matrices and the design matrices for the two approaches as well as other complicated models, to be discussed later. Following the matrices shown in Model 1 of Figure 1, both Andersen's model (Equations (3) and (4)) and Embretson's model (Equations (5) and (6)) are derived.

Five Models for Variations in Item Difficulties across Occasions

Based on variations in item difficulties across occasions and items, five models are proposed. There may be no variations in difficulties across occasions for all items, leading to the *no effect* model, which is equivalent to Andersen's or Embretson's models, in that difficulties are constrained to be unchanged across occasions. The variations may occur in the same amount for all items and occasions, leading to the *constant-occasion/constant-item* model. In addition, they may occur in the same amount for all

		Scoring Matrix		Design Matrix				
It.	*Oc.*	*An.*	*Em.*	*Model* 1	*Model* 2	*Model* 3	*Model* 4	*Model* 5
1	1	100	100	−1 0	−1 0 0	−1 0 0 0	−1 0 0 0 0	−1 0 0 0 0 0
1	2	010	110	−1 0	−1 0−1	−1 0−1 0	−1 0−1 0 0	−1 0−1 0 0 0
1	3	001	111	−1 0	−1 0−1	−1 0 0−1	−1 0−1 0 0	−1 0 0−1 0 0
2	1	100	100	0−1	0−1 0	0−1 0 0	0−1 0 0 0	0−1 0 0 0 0
2	2	010	110	0−1	0−1−1	0−1−1 0	0−1 0−1 0	0−1 0 0−1 0
2	3	001	111	0−1	0−1−1	0−1 0−1	0−1 0−1 0	0−1 0 0 0−1

Note: 1. It.: Item; Oc.: Occasion; An.: Andersen; Em.: Embretson.

2. Model 1: no effect; Model 2: constant-occasion/constant-item; Model 3: variable occasion/constant-item; Model 4: constant-occasion/variable-item; Model 5: variable-occasion/variable-item.

Figure 1. Two matrices for the five models when two item administrated at three occasions based on Andersen's and Embretson's approaches.

items but in different amounts across occasions, leading to the *variable-occasion/constant-item* model. Likewise, they may occur in different amounts for items but the same amount across occasions, leading to the *constant-occasion/variable-item* model. Finally, they may be different for all items across all occasions, leading to the *variable-occasion/variable-item* model.

For the constant-occasion/constant-item model, under Andersen's and Embretson's approaches, the equations are

$$log\left(\frac{P(X_{nik}=1)}{P(X_{nik}=0)}\right)=\theta_{nk}-b_i-d\,,\ d=0 \text{ for } k=1,$$

$$log\left(\frac{P(X_{nik}=1)}{P(X_{nik}=0)}\right)=\sum_{m=1}^{k}\theta_{nm}-b_i-d\,,\ d=0 \text{ for } k=1,$$

respectively. Note that at the first occasion, the variation parameter d is zero because the first occasion is treated as reference. At the following occasions, only the single variation parameter d accounts for all variations for all items.

If a single variation parameter cannot account for all the variations one distinct parameter for each following occasion can be set. For example, two variation parameters are needed for three occasions. The log-odds, $log\,[P(X_{nik}=1)/P(X_{nik}=0)]$, are equal to $\theta_{nk}-b_i-d_{k-1}\,,\ d_{k-1}=0$ for $k=1$,

$$\sum_{m=1}^{k}\theta_{nm}-b_i-d_{k-1}\,,\ d_{k-1}=0 \text{ for } k=1,$$

respectively, for Andersen's and Embretson's, where d_{k-1} represents the variations in difficulties from occasion $k-1$ to occasion k for $k>1$. Note that d_{k-1} does not depend on items, meaning that the variations in difficulties are assumed to be constant for all items. This is why it is referred to as the variable-occasion/constant-item model.

In the constant-occasion/variable-item model, each item has its own variation but these variations are constant across occasions. The log-odds, under Andersen's and Embretson's approaches are equal to

$$\theta_{nk}-b_i-d_i\,,\ d_i=0 \text{ for } k=1,$$

$$\sum_{m=1}^{k}\theta_{nm}-b_i-d_i\,,\ d_i=0 \text{ for } k=1,$$

respectively, where d_i depends on items but not occasions. Although b_i and d_i are in the same equations, they are still identifiable because b_i is "anchored" at the first occasion.

We can go one step further. The variations can vary across not only items but also occasions, which leads to the variable-occasion/variable-item model. In such cases, the log-odds are equal to

$$\theta_{nk} - b_i - d_{i,k-1}, \; d_{i,\,k-1} = 0 \text{ for } k = 1,$$

$$\sum_{m=1}^{k} \theta_{nm} - b_i - d_{i,k-1}, \; d_{i,\,k-1} = 0 \text{ for } k = 1,$$

respectively, for Andersen's and Embretson's approaches, where $d_{i,\,k-1}$ depends on both items and occasions.

Polytomous Cases

All the above modelings focus on dichotomous cases. The MRCML can be easily extended to polytomous cases. The partial credit model, shown in Equation (1), after reparameterization can be expressed as

$$log\left(\frac{P(X_{nij}=1)}{P(X_{nij-1}=1)}\right) = \theta_n - b_{ij},$$

where b_{ij} can be viewed as the difficulty of step j. Note that the log-odds of falling in response j to j - 1 follow the Rasch model. For the rating scale model, the equation becomes

$$log\left(\frac{P(X_{nij}=1)}{P(X_{nij-1}=1)}\right) = \theta_n - b_{i.} - t_j,$$

where $b_{i.}$ is the overall difficulty of item i (the average of all b_{ij}); t_j is the threshold difficulty of step j (the deviation of b_{ij} to $b_{i.}$).

Consider the polytomous items are administrated on K occasions. For the no effect model, the log-odds, $log[P(X_{nijk} = 1)/P(X_{nij\text{-}1k} = 1)]$, of the partial credit model under Andersen's and Embretson's approaches are equal to

$$\theta_{nk} - b_{ij},$$

and

$$\sum_{m=0}^{k} \theta_{nm} - b_{ij},$$

respectively. Likewise, the log-odds for the rating scale modeling are equal to

$$\theta_{nk} - b_{i.} - t_j,$$

$$\sum_{m=0}^{k} \theta_{nm} - b_{i.} - t_j,$$

respectively. For simplicity, we shall not go into details how the MRCML works on these situations. Interested readers may generalize the matrices from those shown in Figure 1. Consider the constant-occasion/constant-item model, a set of parameters is added, one parameter for each step. The log-odds are equal to

$$\theta_{nk} - b_{ij} - d_j, d_j = 0 \text{ for } k = 1,$$

$$\sum_{m=0}^{k} \theta_{nm} - b_{ij} - d_j, d_j = 0 \text{ for } k = 1,$$

respectively for Andersen's and Embretson's. If the d_js are similar, we can further constrain them to be equal, d, resulting in a simpler model. For the rating scale model, the log-odds are equal to

$$\theta_{nk} - b_{i.} - t_j - d, d = 0 \text{ for } k = 1,$$

$$\sum_{m=0}^{k} \theta_{nm} - b_{i.} - t_j - d, d = 0 \text{ for } k = 1.$$

Regarding the variable-occasion/constant-item model, the log-odds are equal to

$$\theta_{nk} - b_{ij} - d_{j,k-1}, d_{j,k-1} = 0 \text{ for } k = 1,$$

$$\sum_{m=0}^{k} \theta_{nm} - b_{ij} - d_{j,k-1}, d_{j,k-1} = 0 \text{ for } k = 1,$$

for the partial credit model, and are equal to

$$\theta_{nk} - b_{i.} - t_j - d_{k-1}, d_{k-1} = 0 \text{ for } k = 1,$$

$$\sum_{m=0}^{k} \theta_{nm} - b_{i.} - t_j - d_{k-1}, d_{k-1} = 0 \text{ for } k = 1,$$

for the rating scale model. For the constant-occasion/variable-item model, the log-odds are equal to

$$\theta_{nk} - b_{ij} - d_{ij}, d_{ij} = 0 \text{ for } k = 1,$$

$$\sum_{m=0}^{k} \theta_{nm} - b_{ij} - d_{ij}, d_{ij} = 0 \text{ for } k = 1,$$

$$\theta_{nk} - b_{i.} - t_j - d_i, d_i = 0 \text{ for } k = 1,$$

$$\sum_{m=0}^{k} \theta_{nm} - b_{i.} - t_j - d_i, d_i = 0 \text{ for } k = 1,$$

respectively. For the most general model, the variable-occasion/variable-item model, the log-odds are equal to

$$\theta_{nk} - b_{ij} - d_{ij,k-1}, d_{ij,\,k-1} = 0 \text{ for } k = 1,$$

$$\sum_{m=0}^{k} \theta_{nm} - b_{ij} - d_{ij,k-1}, d_{ij,\,k-1} = 0 \text{ for } k = 1,$$

$$\theta_{nk} - b_{i.} - t_j - d_{i,k-1}, d_{i,\,k-1} = 0 \text{ for } k = 1,$$

$$\sum_{m=0}^{k} \theta_{nm} - b_{i.} - t_j - d_{i,k-1}, d_{i,\,k-1} = 0 \text{ for } k = 1,$$

respectively.

Table 1 summarizes the various models for dichotomous and polytomous cases under Andersen's and Embretson's approaches. In addition to these five kinds of models, users of the MRCML are able to identify variations in difficulty for some particular items (as to be shown

in simulation studies and real data analyses). With this information, they can gain deeper understanding about these items. Further item revision can be made accordingly.

Simulation Studies

Two simulation studies were conducted. Ten dichotomous items and seven polytomous items were generated in Studies one and two, respectively. The underlying person ability distributions in the studies are multivariate normal. Those generating values for the item parameters are between -1.2 and 1.2 logits for Study one. In Study two, the generating values are the estimates derived from the Family subtest in the following section. Thirty replications were made in each study. Within each study, both Andersen's and Embretson's approaches were used. For simplicity, only two dimensions (or occasions) were generated.

Study One

The testing situation in this study is that ten dichotomous items are administrated in two occasions (e.g., pretest and posttest). We generated items using Andersen's and Embretson's approaches, respectively, and analyzed accordingly. These 10 items follow the Rasch dichotomous model. Because of model identification, the sum of the item parameters was set be to zero, resulting in only 9 item parameters were estimated. The generating item parameters are between -1.2 and 1.2 logits. The sample size is 200.

Under Andersen's approach, the means of the person ability distribution are -.4 and .3, respectively. The variances are .8 and 1.2, respectively, with a covariance of .9. With respect to the recovery of the item parameters, the biases of estimates are very small, ranging from -.019 to .020. With respect to the recovery of the means of the person ability distribution parameters, the biases are even smaller, -.001 and -.006 for the two dimensions, respectively. The biases of the variance-covariance matrix are between -.071 and .020. In sum, all the parameters were recovered very well.

Under Embretson's approach, the first dimension is the initial ability, and the second dimension is the modifiability. The generating means of the person ability distribution means are -.4 and .3, as under Andersen's approach.However, the variances of these two dimensions were 1.0 and .7, respectively. The covariance was .0, meaning that these two dimensions are independent. Regarding the recovery, the biases of the item parameters are quite small, between -.036 and .032. Those for the person ability distribution are small, too, between .044 and -.010. In Embretson's

Table 1

Various Models for Dichotomous and Polytomous Cases Under Andersen's and Embretson's Approaches

Model	Andersen's	Embretson's
Dichotomous Case		
$log\left[P(X_{nik}=1)/P(X_{nik}=0)\right]=$		
1. no effect	$\theta_{nk}-b_i$	$\sum_{m=1}^{k}\theta_{nm}-b_i$
2. constant-occasion/constant-item	$\theta_{nk}-b_i-d$	$\sum_{m=1}^{k}\theta_{nm}-b_i-d$
3. variable-occasion/constant-item	$\theta_{nk}-b_i-d_{k-1}$	$\sum_{m=1}^{k}\theta_{nm}-b_i-d_{k-1}$
4. constant-occasion/variable-item	$\theta_{nk}-b_i-d_i$	$\sum_{m=1}^{k}\theta_{nm}-b_i-d_i$
5. variable-occasion/variable-item	$\theta_{nk}-b_i-d_{i,k-1}$	$\sum_{m=1}^{k}\theta_{nm}-b_i-d_{i,k-1}$
Polytomous Case		
$log\left[P(X_{nijk}=1)/P(X_{nij-1k}=1)\right]=$		
Partial Credit Modeling		
1. no effect	$\theta_{nk}-b_{ij}$	$\sum_{m=0}^{k}\theta_{nm}-b_{ij}$
2. constant-occasion/constant-item	$\theta_{nk}-b_{ij}-d_j$	$\sum_{m=0}^{k}\theta_{nm}-b_{ij}-d_j$
3. variable-occasion/constant-item	$\theta_{nk}-b_{ij}-d_{j,k-1}$	$\sum_{m=0}^{k}\theta_{nm}-b_{ij}-d_{j,k-1}$
4. constant-occasion/variable-item	$\theta_{nk}-b_{ij}-d_{ij}$	$\sum_{m=0}^{k}\theta_{nm}-b_{ij}-d_{ij}$
5. variable-occasion/variable-item	$\theta_{nk}-b_{ij}-d_{ij,k-1}$	$\sum_{m=0}^{k}\theta_{nm}-b_{ij}-d_{ij,k-1}$
Rating Scale Modeling		
1. no effect	$\theta_{nk}-b_{i.}-t_j$	$\sum_{m=0}^{k}\theta_{nm}-b_{i.}-t_j$
2. constant-occasion/constant-item	$\theta_{nk}-b_{i.}-t_j-d$	$\sum_{m=0}^{k}\theta_{nm}-b_{i.}-t_j-d$
3. variable-occasion/constant-item	$\theta_{nk}-b_{i.}-t_j-d_{k-1}$	$\sum_{m=0}^{k}\theta_{nm}-b_{i.}-t_j-d_{k-1}$
4. constant-occasion/variable-item	$\theta_{nk}-b_{i.}-t_j-d_i$	$\sum_{m=0}^{k}\theta_{nm}-b_{i.}-t_j-d_i$
5. variable-occasion/variable-item	$\theta_{nk}-b_{i.}-t_j-d_{i,k-1}$	$\sum_{m=0}^{k}\theta_{nm}-b_{i.}-t_j-d_{i,k-1}$

(1991) simulation studies, a negative bias in the correlation between the initial ability and the modifiability (r = -.210) was found when the generating correlation was zero. In the study here, no substantial bias was found on both the item and the person ability distribution parameters.

Table 2

Summary Result of the Generating Values, Means of Estimates, and Bias of Estimates of 10 Dichotomous Items from 30 Replications Under Embretson's Approach

Generating Values	Means of Estimates	Bias of Estimates
Item Parameter		
-1.2	-1.204	-.004
-1.0	-.979	.020
-.8	-.792	.007
-.6	-.563	.036
-.4	-.430	-.030
.2	.202	.002
.6	.618	.018
.8	.782	-.017
1.2	1.226	.026
Variation Parameter		
-.5	-.560	-.060
.2	.296	-.003
Person Population		
-.2 (Mean of θ_1)	-.211	-.011
.5 (Mean of θ_2)	.506	.006
1.0 (Var. of θ_1)	1.001	.001
.7 (Var. of θ_2)	.668	-.031
.4 (Covariance)	.346	-.053

We also set two items to have variation parameters across occasions under Embretson's approach. As shown in Table 2, all the parameters were recovered quite well.

Study Two

Seven 5-point items following the rating scale model were generated. Like in Study 1, both Andersen's and Embretson's approaches were applied. Thirty replications were conducted with a sample size of 224. The results of Andersen's and Embretson's approaches are summarized in Tables 3 and 4, respectively. All the parameters were recovered very well.

Real Data Analyses

Subjects are 224 college students from Taiwan. They all left homes and live in dormitories, apartments, or their relatives'. The inventory contains thirteen 5-point Likert-type items, ranging from 1 (never) to 5 (always). The first set of seven items concerns the relationship with families (re-

Table 3
Summary Result of the Generating Values, Means of Estimates, and Bias of Estimates of Seven 5-point Scale Items from 30 Replications of the Rating Scale Model Under Andersen's Approach

Generating Values	Means of Estimates	Bias of Estimates
Item Parameter		
-1.497	-1.463	.033
.428	.414	-.013
.920	.912	-.007
1.166	1.153	-.012
-.338	-.338	-.000
-.716	-.711	.004
.294	.255	-.038
.464	.467	.003
-.229	-.234	-.005
Person Population		
-.448 (Mean of θ_1)	-.433	.014
.069 (Mean of θ_2)	.065	-.003
.873 (Var. of θ_1)	.845	-.028
1.699(Var. of θ_2)	1.641	-.057
1.139 (Covariance)	1.066	-.072

ferred to as the Family subtest). The second set of six items concerns the relationship with peers (referred to as the Peer subtest). These two sets can be viewed as subtests, with each subtest taping into a latent trait. Subjects were first asked to recall their relationships with their families and peers before leaving home for college education and then respond to the items. Subsequently, they were asked to respond to the same items based on the relationships at present with their families and peers.

Negative items were recoded before analysis. Consequently, a higher score represents a better relationship. Table 5 shows some descriptive statistics. From the means of the raw scores and the gains of each subtest, we find that both the student-family relationship and the student-peer relationship become better after the move. The two subtests are highly correlated (between .63 and .83). The change of the Family subtest is not correlated with the initial level ($r = -.01$), whereas the change of the Peer subtest is negatively correlated with the initial level ($r = -.36$).

The rating scale modeling was adopted. Each subtest was treated as a unidimensional test and analyzed separately under Andersen's and

Table 4
Summary Result of the Generating values, Means of Estimates, and Bias of Estimates of Seven 5-point Scale Items from 30 Replications of the Rating Scale Model under Embretson's Approach

Generating Values	Means of Estimates	Bias of Estimates
Item Parameter		
-1.399	-1.409	-.010
.414	.417	.003
.912	.901	-.010
1.161	1.149	-.011
-.362	-.368	-.006
-.744	-.733	.010
.326	.327	.001
.459	.461	.002
-.239	-.255	-.016
Person Population		
-.435 (Mean of θ_1)	-.463	-.028
.524 (Mean of θ_2)	.521	-.003
.883 (Var. of θ_1)	.870	-.012
.335 (Var. of θ_2)	.351	.015
.270 (Covariance)	.272	.002

Table 5
Descriptive Statistics Based on the Raw Scores of the Family Subtest, the Peer Subtest, and Change of Each Subtest Before and After the Move

Variable	Mean	Standard Deviation
(1) Family: Before	18.31	6.42
(2) Peer: Before	16.39	5.42
(3) Family: After	21.38	7.78
(4) Peer: After	18.07	5.62
(5) Change of Family: (3) - (1)	3.07	4.49
(6) Change of Peer: (4) - (2)	1.68	4.40

Correlations

	(1)	(2)	(3)	(4)	(5)
(2)	.66				
(3)	.82	.63			
(4)	.75	.68	.83		
(5)	-.01	.15	.56	.37	
(6)	.16	-.36	.29	.44	.29

Embretson's approaches. The estimates of the parameters of the two subtest are listed in Table 6. Regarding the item parameters, those based on Andersen's approach are almost identical to those on Embretson's. This is naturally because they are all based on the same sufficient statistics. Re-

Table 6
Item Parameter Estimates and Standard Errors
Based on the Consecutive Method

	Family				Peer			
	Andersen's		Embretson's		Andersen's		Embretson's	
	Estimate	S.D.	Estimate	S.D.	Estimate	S.D.	Estimate	S.D.
Overall Difficulty								
1	1.50	.07	1.40	.06	-1.78	.07	-1.78	.07
2	-.43	.05	-.41	.05	-.34	.05	-.34	.05
3	-.92	.05	-.91	.05	.84	.05	.83	.05
4	-1.17	.05	-1.16	.05	.24	.05	.24	.05
5	.34	.05	.36	.05	-.49	.05	-.49	.05
6	.72	.05	.74	.05	1.53		1.54	
7	-.04		-.02					
Threshold Difficulty								
1	-.30	.06	-.33	.06	-.81	.06	-.81	.06
2	-.46	.07	-.46	.07	-.07	.07	-.07	.07
3	.23	.07	.24	.07	.52	.04	.52	.08
4	.53		.55		.36		.36	
Person Population								
Mean (θ_1)	-.45		-.44		-.29		-.29	
Mean (θ_2)	.07		.52		.06		.36	
Var. (θ_1)	.88		.88		1.09		1.08	
Var. (θ_2)	1.70		.34		1.12		.41	
Covariance	1.14		.27		.90		-.18	
Likelihood Statistic								
	7401.89		7405.99		6442.38		6442.33	

garding the person ability distribution parameters, one can apply Equations (8) and (9) to verify the relationship between these two approaches.

In both the Family and the Peer subtests, the means of the modifiability are positive, meaning that both the student-family and the student-peer relationships became better after the students left homes. This can be also found under Andersen's approach, because the means of the second dimension are somewhat higher than those of the first dimension.

The correlations between the two dimensions for the Family and the Peer subtests are .93 and .81, respectively, under Andersen's approach. This is as expected because these items were administrated twice. However, those correlations are .49 and -.27, respectively, under Embretson's, meaning the modifiability is positively correlated with the initial ability in the Family subtest, and negatively correlated with the initial ability in the Peer subtest. Therefore, students with high initial levels in the Family subtest dimension (i.e., better student-family relationship before the move) become higher after the move in an accelerated rate than those with low

initial levels. In contract, the rate is descending in the Peer subtest, although the relationship still gets better.

In the above analyses, the two subtests were analyzed separately, which is referred to as the *consecutive* method. In fact, the two subtests are correlated, the performances on one subtest can provide some collateral information on the other. The whole test, with two unidimensional subtests, can be viewed as a *multidimensional between-item* test (Wang, Wilson, & Adams, 1997), because each item itself is unidimensional but a set of items express multidimensionality. The MRCML is flexible to allow the two subtests to be analyzed simultaneously. We also applied this *simultaneous* method together with Embretson's approach an. The estimates and their standard errors are shown in Table 7. The estimates of the item parameters and the means of the person ability distribution are quite close to those based on the consecutive method.

Regarding the model fit, the model based on the simultaneous method yields a loglikelihood statistic of 13675.38 with 31 parameters estimated. The two loglikelihood statistics based on the consecutive method are 7405.99 and 6442.33 with 14 and 13 parameters estimated, respectively, for the Family and the Peer subtests. The sum of the two loglikelihood statistics are 13848.32 with 27 parameters estimated. To compare these two methods, we adopt Akaike's information criterion (*AIC*; Akaike, 1977). In terms of the *AIC*s, the model based on the simultaneous method, with an *AIC* of 13737.38, fits the data better than those based on the consecutive method, with an *AIC* of 13902.32.

Variations in Item Difficulty across Occasions

The above analyses were based on the no effect model. We have applied more complicated models to investigate variations in item difficulties across occasions. Since there are only two occasions, the variable-occasion/constant-item model is equivalent to the constant-occasion/constant-item model. They are both referred to as the constant-item models. Likewise, the variable-occasion/variable-item model is equivalent to the constant-occasion/variable-item model, and both are referred to as the variable-item models.

Table 8 lists the likelihood statistics and the number of parameters for the no effect, the constant-item, and the variable-item models. Since the three models are nested, the usual likelihood ratio test was applied for model comparisons. The variable-item model is significantly better than

Table 7
Item Parameter Estimates and Standard Errors based on the Simultaneous Method under Embretson's Approach

	Family		Peer	
Overall Difficulty	Estimate	S.D.	Estimate	S.D.
1	1.57	.07	-1.97	.08
2	-.46	.05	-.38	.05
3	-1.03	.05	.94	.05
4	-1.30	.06	.27	.05
5	.41	.05	-.50	.05
6	.83	.05	1.64	
7	-.02			
Threshold Difficulty				
1	-.48	.06	-1.09	.06
2	-.51	.07	.05	.08
3	.28	.07	.55	.08
4	.71		.49	
Person Population				
Mean ($\theta_{1,\text{Family}}$)	-.47			
Mean ($\theta_{2,\text{Family}}$)	.56			
Mean ($\theta_{3,\text{Peer}}$)	-.28			
Mean ($\theta_{4,\text{Peer}}$)	.38			
Covariance-Variance Matrix				
	θ_1	θ_2	θ_3	θ_4
θ_1	1.26			
θ_2	.16	.65		
θ_3	.91	.28	1.32	
θ_4	.20	.28	-.19	.68
Likelihood Statistic	13675.38			

Note: θ_1: Initial ability of the Family subtest; θ_2: Modifiability ability of the Family subtest; θ_3: Initial ability of the Peer subtest; θ_1: Modifiability ability of the Peer subtest.

the other two models, in terms of model fit. Table 9 lists the variation parameters and their standard errors for the variable-item model. They are not very far away from zero, except those for item 3 in both the Family and Peer subtests. According to the signs of these two variation parameters, both items become more difficult at the second occasion.

Since the other variation parameters are close to zero, perhaps only these two items express the variations. To investigate this, we conducted another model, the *partial-item* model, where only two variation parameters

were estimated. The estimates of these two variation parameters are .20 and .19, respectively, both with a standard error of .05. This partial-item model is not significantly different from the variable-item model ($\Delta G^2 = 16.73$, $\Delta df = 11$, $p > .05$). Therefore, only these two items express the variations in difficulties across occasions and thus call for further investigation.

Table 8
Number of Parameters and Likelihood Statistics of the Four Models

Model	No. of Parameters	Likelihood Statistic
no effect	31	13675.38
constant-item	33	13672.55
variable-item	44	13641.73
partial-item	33	13658.46

Table 9
Variation Parameter Estimates and Standard Errors of the Variable-Item Model Based on the Simultaneous Method under Embretson's Approach

	Estimates	Standard Errors
Family Subtest		
1	-.20	.12
2	.02	.09
3	.39	.10
4	-.15	.10
5	-.09	.10
6	-.08	.10
7	.13	.10
Peer Subtest		
1	-.06	.11
2	-.15	.09
3	.36	.10
4	.04	.10
5	-.23	.09
6	.14	.10

Conclusions

Change measurement has been a major issue within classical test theory. Recent developments in IRT have made it applicable to this issue. Fischer and his colleagues have proposed a serial of models for treatment and trend effects in the measurement of change. These models, factoring out the person ability distribution in estimating of the item parameters, do not

directly estimate the individual differences in change. Andersen's model, although including impact of time or treatment on the ability distribution, does not directly estimate the individual differences in change. Embretson's model, building in individual the change parameters, provides estimates of the individual difference in change.

Despite the progress in measuring individual differences in change within IRT, several practical issues yet to be addressed. First, both Andersen's and Embretson's models are limited to dichotomous items. Second, these two models, assuming item difficulties remain constant across occasions or conditions, may be too strict because items may express different difficulties when administrated repeatedly, due to practice, memory, or response consistency effects. This kind of variations in item difficulties should be checked. In this article, we present a newly developed multidimensional Rasch measurement model, the MRCML. It is characterized by a scoring matrix and a design matrix. Manipulating these two matrices, we show how both Andersen's and Embretson's approaches are extended to polytomous items and to investigation of variations in item difficulties across occasions.

Specifically, based on variations in difficulties across occasions and items, five kinds of models are categorized. In the no effect model, item difficulties remain unchanged across occasions, which is equivalent to Andersen's and Embretson's approaches extended to polytomous items. In the constant-occasion/constant-item model, only one variation parameter is added to represent the variations for all items across all occasions. In the variable-occasion/constant-item model, one variation parameter is added for each following occasion. In the constant-occasion/variable-item model, one variation parameter is added for each item. In the variable-occasion/variable-item model, one parameter is added for each item at each following occasion. Users are also able to add variation parameters for only a subset of items. Through model comparisons, deeper understanding of the variations in difficulties across occasions can be gained.

In the simulation studies, we adopted both Andersen's and Embretson's approaches to dichotomous items and polytomous items. Some items expressing the variations were studied, too. No substantial bias was found. In the real data analyses, we followed Andersen's and Embretson's approaches and imposed a rating scale model. Each subtest was analyzed not only consecutively but also simultaneously. The simultaneous method is better than the consecutive method in terms of model fit. In addition,

one item in each subtest was found expressing variation in difficulties. They call for further investigation and revision.

References

Adams, R. J., & Wilson, M. R. (1996). Formulating the Rasch models as a mixed coefficients multinomial logit. In G. Engelhard & M. R. Wilson (Eds.), *Objective measurement: Theory into practice.* Vol. III. Norwood, NJ: Ablex.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

Akaike, H. (1977). On entropy maximization principle. In P. R. Krischnaiah (Ed.), *Applications of statistics.* New York: North Holland.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50,* 3-16.

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

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.). *Problems in measuring change.* Madison: University of Wisconsin Press.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression / correlation for the behavior sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change" - or should we? *Psychological Bulletin, 74,* 68-80.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495-515.

Fischer, G. H., & Pazer, P. (1991). An extension of the rating scale model with an application to the measurement of change. *Psychometrika, 56,* 637-651.

Fischer, G. H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika, 59,* 177-192.

Geenen, R., & van de Vijver, F. J. R. (1993). A simple test of the Law of Initial Values. *Psychophysiology, 30,* 525-530.

Gottman, J. H., & Krokoff, L. J. (1990). Complex statistics are not always clear than simple statistics: A reply to Woody and Costanzo. *Journal of Consulting and Clinical Psychology, 58,* 502-505.

Jamieson, J. (1993). The Law of Initial Values: Five factors or two? *International Journal of Psychophysiology, 14,* 233-239.

Jamieson, J. (1994). Measurement of change and the Law of Initial Values: A computer simulation study. *Educational and Psychological measurement, 55,* 38-46.

Jin, P. (1992). Toward a reconceptualization of the Law of Initial Value. *Psychological Bulletin, 111,* 176-184.

Llabre, M. M., Spitzer, S. S., Saab, P. G., Ironson, G. H., Schneiderman, N. (1991). The reliability and specificity of delta versus residualized change as measures of cardiovascular reactivity to behavioral challenges. *Psychophysiology, 28,* 701-711.

Lord, F. E. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change.* Madison: University of Wisconsin Press.

Malgady, R. G., & Colon-Malgady, G. (1991). Comparing the reliability of difference scores and residuals in analysis of covariance. *Educational and Psychological Measurement, 51,* 803-807.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Rasch, G. (1960 / 1980). *Probabilistic Models for Some Intelligent and Attainment Tests.* Copenhagen: Danmarks Paedogogiske Institut.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 5,* 321-333.

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin, 109,* 147-151.

Wang, W., Wilson, M. R., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard & K. Draney (Eds.), *Objective measurement: Theory into practice.* Volume 4 (pp. 139-155). Norwood, NJ: Ablex.

JOURNAL OF OUTCOME MEASUREMENT®, 2(3), 266 - 283

Detecting Multidimensionality: Which Residual Data-type Works Best?

John Michael Linacre
MESA Psychometric Laboratory
University of Chicago

Factor analysis is a powerful technique for investigating multidimensionality in observational data, but it fails to construct interval measures. Rasch analysis constructs interval measures, but only indirectly flags the presence of multidimensional structures. Simulation studies indicate that, for responses to complete tests, construction of Rasch measures from the observational data, followed by principal components factor analysis of Rasch residuals, provides an effective means of identifying multidimensionality. The most diagnostically useful residual form was found to be the standardized residual. The multidimensional structure of the Functional Independence Measure (FIM℠) is confirmed by means of Rasch analysis followed by factor analysis of standardized residuals.

Requests for reprints should be sent to John Michael Linacre, MESA Psychometric Laboratory, University of Chicago, 5835 S. Kimbark Ave., Chicago, IL 60637

Introduction

The Rasch model constructs a one-dimensional measurement system from ordinal data, regardless of the dimensionality of those data. Empirical data are always manifestations of more than one latent dimension. For instance, in observational instruments, the observer's own training level and perspectives influence the observations recorded. In self-administered tests, the ability of the subject to comprehend and follow instructions becomes part of the subject's self-assessment. Consequently the Rasch dimension is a composite based on the conjoint ordering of persons, items and other facets of measurement according to their raw scores (with allowance for incomplete data).

When the data accord exactly with the Rasch model, then all systematic variation within the data is explained by the one dimension. The removal of the implications of this dimension (for both persons and items) from the data leaves behind observation-level residuals with a random normal structure and predictable variance (Wright & Masters, 1982, p. 98). Consequently, the residuals for pairs of items across persons are uncorrelated, a property known as "local independence" (Lazarsfeld, 1958). Since Lazarsfeld introduced the term "local independence" in the context of latent class analysis, he conceptualized all relevant persons to be located at the same point on the variable. In Rasch usage and in this paper, local independence is modeled to hold not just for the classes, i.e., at particular points along the variable, but at every point along the variable. Thus local independence is modeled to hold not just at the class level, but for each person. To verify local independence under Rasch model conditions, for which replication of observations is necessary, coincidence of person locations on the latent variable is achieved by removing the effect of different person measures from the observations (Andrich, 1991).

In practice, however, it is impossible to discern, from the data alone, whether a particular residual is an accidental outcome of a process that accords with the Rasch model, or is produced by unmodeled dimensions. Indeed, all deviation in the data from the Rasch dimension could be considered symptoms of multidimensionality. Is an unexpected correct answer on a test the one-in-a-thousand occurrence predicted by the Rasch model, or is it a lucky guess? Even a single random lucky guess on a certification test results in data that confounds a competence dimension with a guessing dimension, causing the Rasch dimension to be a composite of the two. Since the certification information in the data overwhelms

the guessing information, most users are content to label the test a "certification" test, and the Rasch dimension, a "certification" dimension.

A few unusual responses slightly bias the measures toward the center of the test (Adams and Wright, 1994). They also slightly reduce the statistical validity of the measures of the relevant persons and items (Wright and Stone, 1979, pp. 181-190). When such observations are a cause for concern, they can be identified and diagnosed by examining the patterns of responses by the relevant persons or to the relevant items. Since such detailed examination of all the data is unreasonable, it is useful to perform an initial screening of the data using person- and item-level quality control fit statistics, such as Outfit and Infit (Wright and Stone, 1979, pp. 66-82). Gross non-normality of residuals would be detected at this stage.

A pervasive, but usually less obvious, perturbation of the residuals is symptomatic of the presence of more than one dimension in the data. Extra dimensions may reflect different person response styles or different item content areas. Since, unidimensionality is always provisional, and ultimately utilitarian, the occurrence of multiple dimensions in the data does not necessarily imply substantive multi-dimensionality. Certification tests contain both theory and practice aspects, but the data can express a unidimensional "competency" variable.

Multidimensionality only becomes a real concern when there are response patterns in the data indicating that the data represent two or more dimensions so disparate that it is no longer clear what latent dimension the Rasch dimension operationalizes. A data-set manifests one dimension so long as it is productive to think of it that way. For educational policy-makers, math is everything from addition to calculus. For cognitive psychologists, the mental processes underlying addition may be very different from those underlying subtraction.

In the extreme, every test item defines its own dimension. For instance, a common one-item test is the question, "What is your age"? In diagnostic testing, each response to each item may indicate a specific course of action. Nevertheless, the inferential goal is to generalize across as many different items as possible that usefully manifest the same variable, such as "patient independence". Utility is defeated, however, when different subsets of such items would lead to different generalizations. In this instance, utility dictates that what was considered to be the "same" variable is, in fact, two (or more) different variables, each leading to different inferences. An example is the Functional Independence Measure,

FIM[SM]. Though originally intended to generalize one dimension of functional independence across a mixture of 18 motor and cognitive items, closer inspection indicated that it would generally be more useful to use the FIM items to construct separate "motor" and "cognitive" measures for each patient (Linacre et al., 1994).

Multidimensionality can also be an artifact of test construction. For instance, including the identical item several times in a certification test produces a subset of responses to those items that have high inter-correlation across persons. These items define their own idiosyncratic local dimension based on that one item. On the other hand, the use of different response mechanisms across items (multiple-choice, open-ended, rating scales) introduces unmodeled variation in the response-level data that can be attributed to a dimension of "item type" (Wilson and Wang, 1995).

Identifying Statistical Multi-Dimensionality

Since the only multidimensionality of real measurement concern is manifested by unmodeled behavior in the data, it is that part of the data that must be examined. After the construction of Rasch measures from the current data (or their imputation from previous data or by theory), an expected value can be computed for each ordinal observation. The observation residual is the observation less its expectation. It is by looking for patterns among these residuals that relevant multidimensionality can be identified. "Analysis of the fit of data to [local independence] is the statistical device by which data are evaluated for their measurement potential - for their measurement validity" (Wright 1995).

Since there are many ways in which data can depart from the Rasch model (Glas and Verhelst, 1995), it has been suggested that the most blatant departures be investigated first, followed by more subtle ones. Using a comparative example, Linacre (1992) suggested a three stage procedure: (i) remediate systematic contradictions to the Rasch dimension, typically flagged by negative point-biserial correlations; (ii) diagnose idiosyncratic persons and items using local quality-control fit statistics, such as INFIT and OUTFIT; (iii) look for multidimensionality.

It is the residual inter-correlations across items that indicate whether subgroups of items cluster together in a non-homogeneous way, symptomatic of multidimensionality. "The misfit of the Rasch model to a data set can be measured by the size of residual covariances. Unfortunately, some computer programs for fitting the Rasch model do not give any

information about this. A choice would be to examine the covariance matrix of the item residuals, not the sizes of the residuals themselves, to see if the items are indeed conditionally uncorrelated, as required by the principle of local independence" (McDonald, 1985, p. 212).

Conditionally correlated item residuals indicate the presence of other measurement dimensions, beyond the primary dimension. This suggests a two-step process. First, identify the other dimensions. Second, decide whether they are of sufficient interest to warrant the construction of separate measures for those dimensions. This paper focuses on the first step, the identification of secondary dimensions. In this endeavor, principal components factor analysis is used to detect structure in the inter-item residual correlation matrix.

The use of factor analysis to identify the primary dimension in data is discussed by Wright (1996) and Smith (1996). In essence, factor analysis aids in the classification of items into potential dimensions, and assists with the partitioning of raw scores according to those dimensions. It does not however, construct linear measures from the data along those dimensions. Consequently, factor scores and loadings have an uncertain sample dependency and analyst-choice-dependent nature that renders their direct use in subsequent analyses precarious.

Choice of Residual Form for Item Correlations

Consider a simple polytomous form of the Rasch model:

$$\log\left(\frac{P_{nik}}{P_{ni(k-1)}}\right) = B_n - D_i - F_k$$

where P_{nik} is the probability of being observed in ordered category k for person n on item i, where k ranges from 1 to m,

$P_{ni(k-1)}$ is the probability of being observed in category $k-1$ for person n on item i,

B_n is the ability of person n,

D_i is the difficulty of item i, and

F_k is the step difficulty of category k relative to category $k-1$.

Each data-point, X_{ni}, is an observed category in the range 0 to m, resulting from an interaction between person n and item i. Corresponding to each X_{ni} is an expected score, E_{ni}, given by

$$E_{ni} = \sum_{k=0}^{m} k P_{nik}$$

with model variance of the observed outcome about the expected, V_{ni}, where

$$V_{ni} = \sum_{k=0}^{m} (k - E_{ni})^2 P_{nik} = \sum_{k=0}^{m} k^2 P_{nik} - E_{ni}^2$$

This suggests a variety of residuals for investigation regarding inter-item correlation (see Table 1). The raw score residual, Y_{ni}, is the difference between the observed and expected category values and has the range $-m$ to m. Each standardized residual, Z_{ni}, is normalized by its local modeled standard deviation. These standardized residuals are expected to approximate a N(0,1) distribution (Smith, 1988). The logit residual, L_{ni}, is a first approximation to the measurement discrepancy indicated by the raw score residual. The modeled observation variance, V_{ni}, is the raw-score-to-logit conversion factor (Wright and Masters, 1982, p. 77). The relationship between the three residuals can be complex, and depends on the shape of the item information function, defined by the rating scale structure. For a two-category rating scale, i.e., for dichotomous observations, the relationship is shown in Figure 1.

Table 1

Observation Residuals

Residual type	Mathematical expression
Raw score residual	$Y_{ni} = X_{ni} - E_{ni}$
Standardized residual	$Z_{ni} = (X_{ni} - E_{ni}) / (V_{ni})^{1/2}$
Logit residual	$L_{ni} = (X_{ni} - E_{ni}) / V_{ni}$

The choice of which type of residual to employ in the investigation of multidimensionality is not clear cut. A *prima facie* case could be made for each one of them. Since Rasch analysis is a measurement-based approach, investigation of residuals from a measurement-based perspective would appear most productive. This would focus on the logit residuals. On the other hand, since unmodeled patterns in the residuals contradict

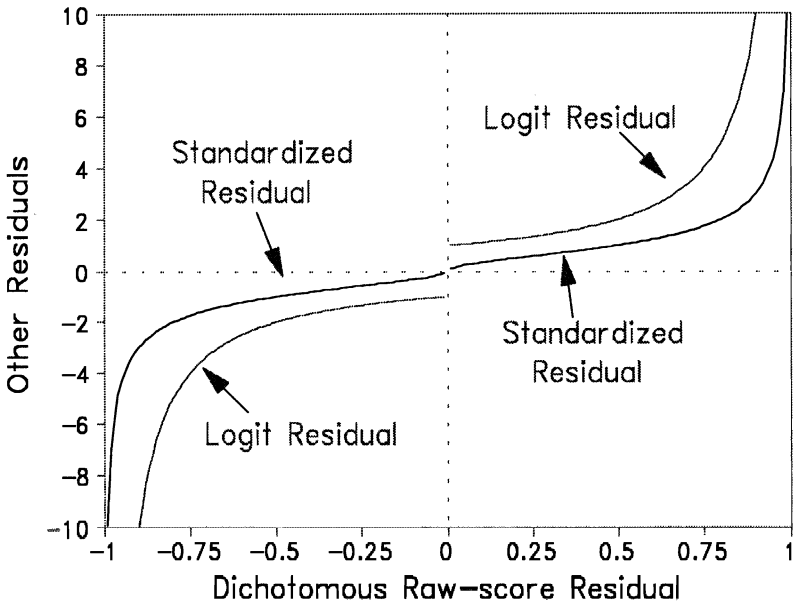

Figure 1. Relationship between residuals to dichotomous observations

the measurement framework, the standardized residuals may have more diagnostic power due to their clear statistical properties. The raw score residuals, however, most directly reflect the presence of any other dimensions. Indeed, these last most closely resemble the original raw observations which are widely used in the investigation of multidimensionality (Thurstone, 1932).

A Simulation Study for Two Dimensions

In view of the uncertainty in the choice of residual with which to compute inter-item correlations, a series of simulation studies was conducted. The purpose of the studies was to discover which form of residual most clearly identified the multidimensional structure underlying the data in straight-forward situations. Principal components analysis, also called the principal-factor method, was chosen for this investigation because of its "rigorous mathematical basis" (Harman, 1960, p. 154). Substitution of common-factor methods in these simulation studies (not reported here) was found to lead to the same conclusions. The simulation studies employ dichotomous items, but the utility of their result is illus-

trated with a polytomous empirical data set.

For the first study, a sample of 1190 persons was generated. Each person was assigned two orthogonal abilities: a "math" ability randomly from an N(0,1.5) logit distribution, and a "reading" ability randomly from a distribution with the same shape. The two abilities were assigned independently, producing orthogonality. A two-dimensional test was then posited containing 3 types of dichotomous items: (i) 100 "math" items uniformly distributed in difficulty over -2 to +2 logits; (ii) 25 "reading" items uniformly distributed over -2 to +2 logits; (iii) 50 "word problem" items (conceptually combine reading and math) uniformly distributed over -2 to +2 logits.

Dichotomous observational data were generated for each person. For the math items, the math ability was used. For the reading items, the reading ability was used. For the word problem items, the *lower* of each person's math and reading abilities was used.

Rasch analysis of this observational data was performed. One measure was estimated for each person across all items and one difficulty for each item across all persons using the BIGSTEPS Rasch analysis program (Wright and Linacre, 1997). Based on these estimated measures, expected observations were obtained and the three score residuals calculated.

Because there are more math than reading items, the primary "Rasch" dimension is expected to be dominated by the math items. The reading items should give the strongest indication of a second dimension. The word problems should cluster halfway between the math and reading items. Smith and Miao (1994) reported that the ratio of 4 items on one dimension to 1 item on another generally produces a dimensional structure that can be identified directly by principal components analysis of the observations themselves. Accordingly, this was done.

Figure 2 shows a plot of the loadings of the first principal component (unrotated) in the simulated data against the Rasch item difficulties estimated from that same data. (Computations were performed by the author using proprietary software which had been validated against standard data sets). The item difficulties fall mainly within their simulated range of -2 to +2 logits. The 0.5 logit increase in the difficulty of the "W" items (word problems) relative to their generators is due to the choice of the lower of math and reading ability in generating the observations. This choice has had the expected effect of making the estimated items appear more difficult than the generating items.

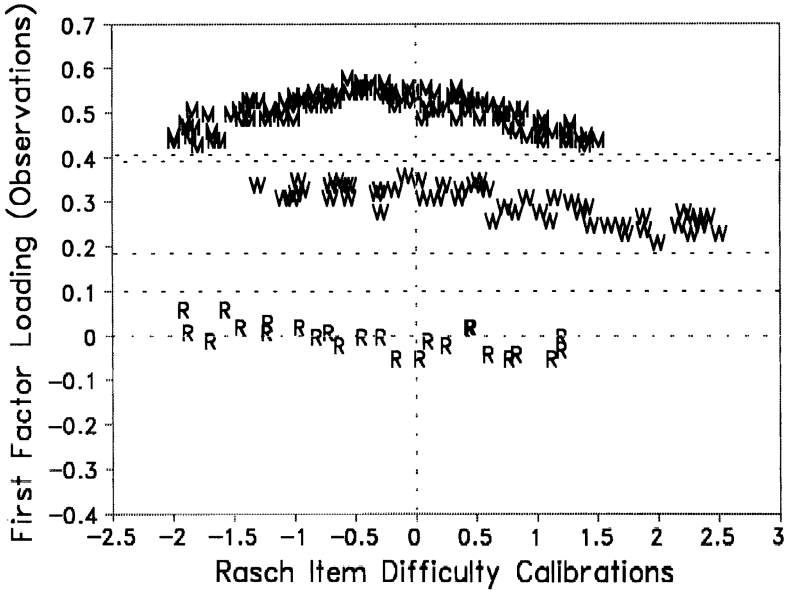

Figure 2. Plot of first principal component of item observations against Rasch item difficulty calibrations.

The loadings on the first principal factor in the observations stratify the items by type: M for math items, W for word problems and R for reading items. The math items show the highest loading on the first factor, the reading items the least, as expected. The effect of item difficulty level is secondary, but the convex form of the "M" distribution indicates that extreme item easiness or difficulty attenuates the loading on the first factor. The non-linearity of raw scores is distorting the factor structure. Consequently, the vertical difference between the lowest M and highest W is small, meaning that the stratification, which is obvious in the plot, would be less striking in a table of factor loadings.

Figure 3 is based on the raw residuals. The factor loadings for the first principal component in the item residual correlations are plotted against item difficulties. The first (Rasch) dimension has been explicitly removed. The highest loadings on this second, residual, dimension are now obtained by the reading items. (Since factor direction is arbitrary,

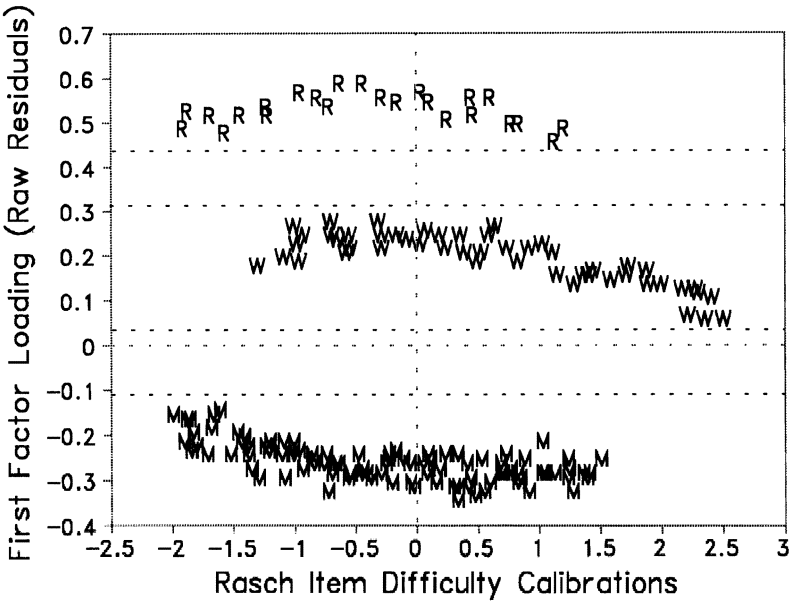

Figure 3. Loadings on first principal component in raw residual correlations against Rasch item calibrations.

the largest factor loading is shown as positive in this study). The fact that the Rasch dimension is a compromise between the math and reading items is confirmed by the negative, rather than zero, loadings of the math items.

The raw residuals produce a better stratified and less curved plot than the original observations. This could have been expected because the data were simulated to fit the Rasch model. Nevertheless, it is encouraging that introducing another orthogonal dimension into the data has not invalidated a Rasch-based dimensional structure.

Figure 4 is based on the standardized residuals. With these data, the differences between the standardized and raw residual plots are barely distinguishable by eye.

Figure 5 employs the logit residuals. This plots shows attenuated loadings on the extreme items, clouding the nature of the dimensionality in the simulated data. Nevertheless, this Figure remains clearer than that based on the observations themselves, Figure 2.

These simulations of dichotomous observations suggest that none of these four approaches would be misleading, but that raw and standardized residuals give the clearest results.

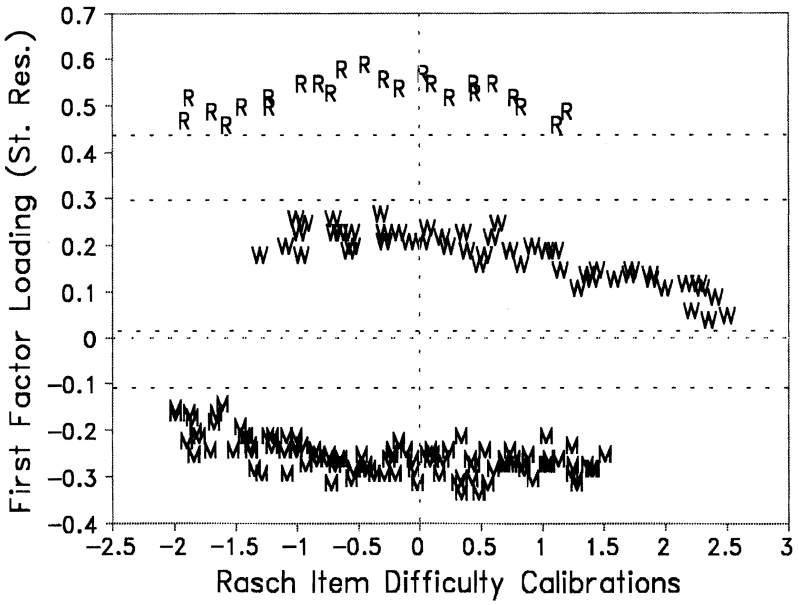

Figure 4. Loadings on first principal component in standardized residual correlations against Rasch item calibrations.

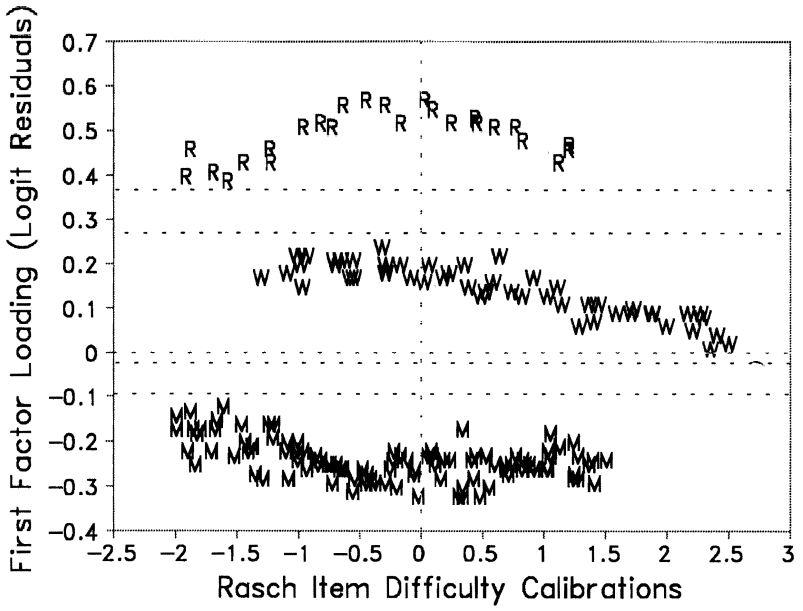

Figure 5. Loadings on first principal component in logit residual correlations against Rasch item calibrations.

Simulation Study: Correlated Dimensions

A more subtle form of multidimensionality is that of correlated dimensions. As a trainee advances through a course of study, both knowledge of theory and practical skills tend to improve, but not exactly in step. This can lead to the trainee having a "knowledge" ability and a "skill" ability. Across a sample of trainees at different stages of their training these abilities will be correlated, but different. A test consisting of both knowledge and skill items will probe both abilities, and the reported trainee measure will be a composite of the two abilities. Analysis of residuals can alert the analyst that this has occurred.

In the second simulation, a sample of 1000 persons was generated. Each person was assigned two abilities: an "X" ability randomly from an N(0,1) logit distribution, and a "O" ability randomly from a distribution with the same shape, but such that the X and O abilities have a 0.9 correlation across the sample. Responses by this sample to a test of 50 X-type and 50 O-type items were simulated, such that each person is modeled to respond to each item type with the corresponding ability, e.g., responses to X items are with X abilities. For each item type, the item difficulties were uniformly distributed from -2.0 to +2.0 logits.

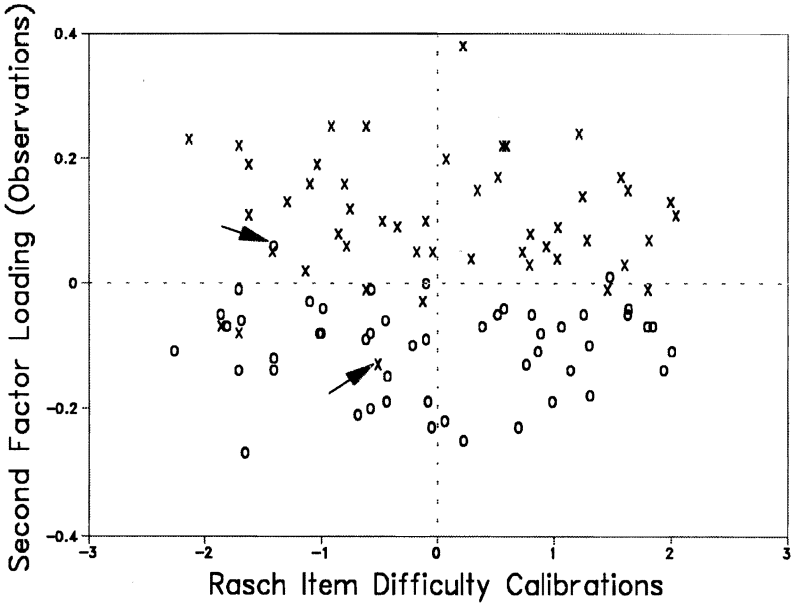

Figure 6. Loadings on the second factor for observations with the correlated multidimensional data.

The inter-ability correlation of 0.9 was set high so that neither principal components factor analysis of the observations nor item-level OUTFIT statistics would be expected to detect the dimensional nature of the items successfully (Smith and Miao, 1994). As a further complication, the mean ability of the sample was set at the center of the test, removing any skewing of the observation variance.

Principal components factor analysis of inter-item correlations was performed. Figure 6 shows the loadings on the *second* factor for these simulated observations. This factor is generally successful in discriminating X and O-type items. The most displaced X and O items are indicated with arrows.

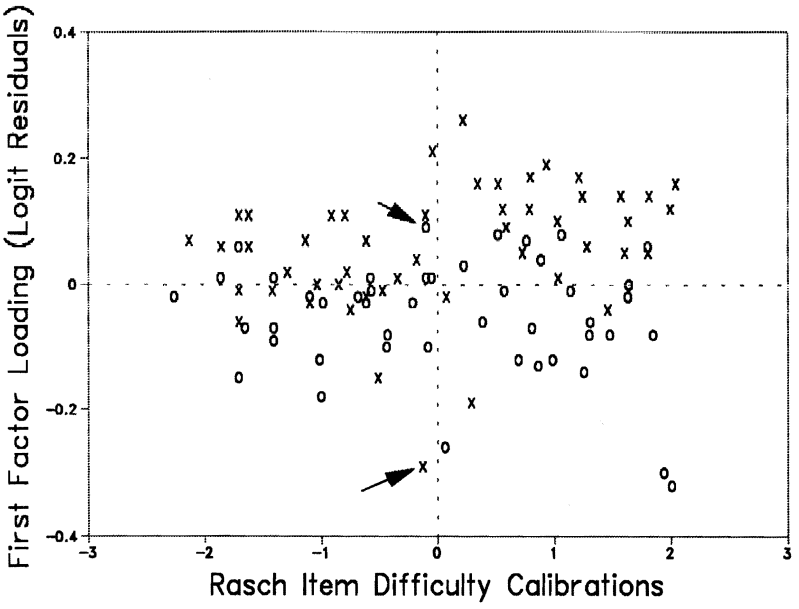

Figure 7. Loadings on the first factor for logit residuals with correlated multidimensional data.

Figure 7 shows the loadings on the first factor for the logit residuals. This approach is less successful in discriminating X and O items. In particular, the most displaced X item, at the bottom of the plot, is indicated to be more strongly O-type than nearly all O items.

Figure 8 shows the loadings on the first factor for the raw residuals. This approach is more successful. Only one O item and one X item are noticeably displaced.

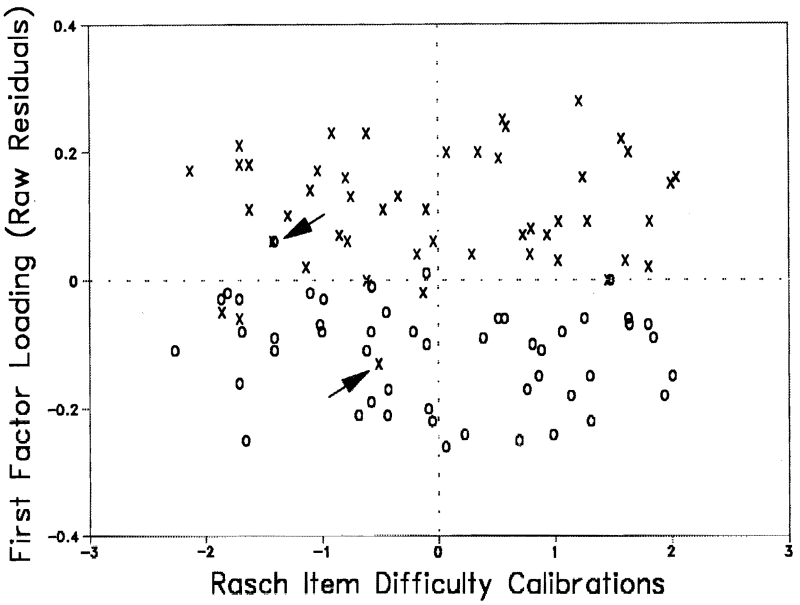

Figure 8. Loadings on first factor for raw residuals with the correlated multidimensional data.

Figure 9 shows the loadings on the first factor for the standardized residuals. This approach is the most successful. Only one X item is noticeably displaced.

In similar simulations, not reported here, but with lower inter-dimensional correlations and different sample-test targeting, this pattern continued. The logit residuals were the least successful at discriminating X and O type items. Factor analysis of the observations themselves was more successful in discriminating item types, but the raw and standardized residuals were most successful and about equally effective.

An Example Application

In order to verify the effectiveness of principal components factor analysis of residuals, Rasch analysis was performed on a random sample of 6,144 FIM[SM] records (from the UDS database, courtesy of Carl V. Granger). Only data collected at the *admission* time point were analyzed. Figure 10 plots the loadings of the first factor in the standardized residuals against the logit calibrations of the 18 FIM items. This Figure immediately signals the divergence of the five cognitively-oriented items (top of the Figure) from the thirteen motor-oriented items. This same diver-

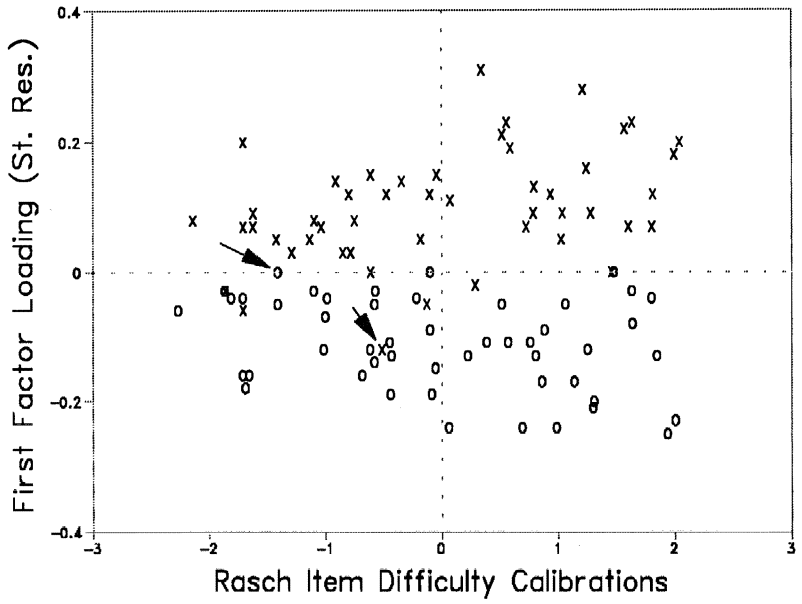

Figure 9. Loadings on the first factor for standardized residuals with the correlated multidimensional data.

gence was reported in Linacre et al. (1994), but only after a tortuous analysis of admission and discharge data. For these FIM data, the raw residual plot was almost identical to Figure 10, but with slightly less range to the loadings. Both identify the opposite poles of the factor to be "memory" and "toilet transfer". Analyses of the logit residuals and the original observations each generated a minor factor that corresponded to the cognitive-motor contrast, but with different orderings of the items at each end of the factor. For the raw observations, the extremes are "comprehension" vs. "stairs". For the logit residuals, "bathing" vs. "problem solving". Thus, though the standardized residuals provided the distinct solution, the clinical implications of the factor structure might direct the analyst to favor use of a different residual for this analysis.

Once divergence within an item pool has been identified, the next step is to evaluate its impact on measurement. For the FIM, this is investigated by measuring the sample, first on the variable defined by the five cognitive items, then on that defined by the thirteen motor items. When the differences between the resulting pairs of measures have clinical implications, e.g., when one measure indicates normal functioning and the other dysfunction, the multidimensionality of the original instrument is

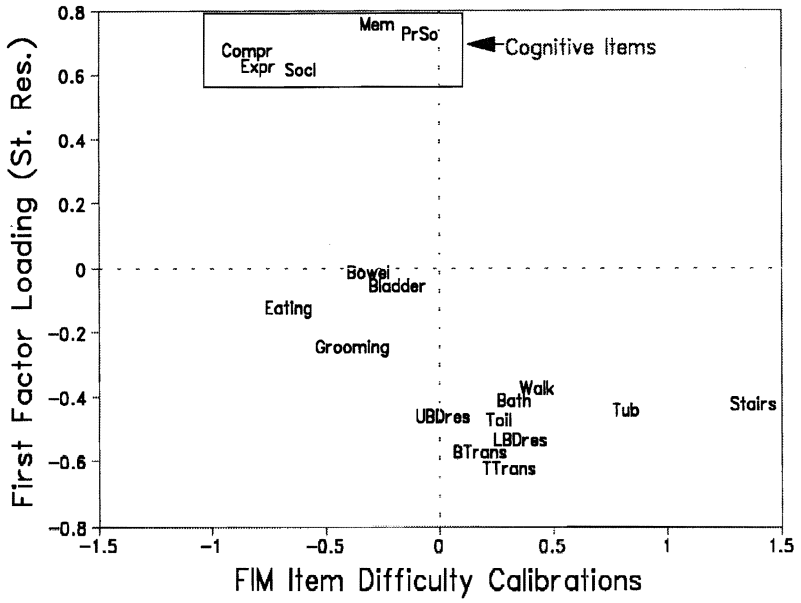

Figure 10. Loadings on the first factor for standardized residuals with the FIM data.

resolved by setting up two measurement systems. This is the case for many applications of the FIM. When differences between the pairs of measures have no implications for practice, then the multidimensionality is treated as an unwanted, but inevitable, source of the noise within the data, slightly lowering the quality of the one measurement system.

Conclusion

For complete tests, principal components factor analysis of either the observations themselves or the various residual formulations successfully reflects the multidimensional structures simulated here. Though these simulated structures are more clear-cut than those hypothesized to exist in empirical data, the essential features are likely to encompass the same structures: items of varying dimensionality and persons with multiple, but correlated, abilities. A word of caution: empirical data often incorporate departures from the Rasch model that would distort the distribution of the residuals, including miskeyed items, data entry errors and response sets. It is recommended that these issues be addressed prior to factor analysis.

Overall, standardized residuals provided the most decisive analysis,

but their advantage over raw residuals was slight. Logit residuals were less informative.

Factor analysis of the observations themselves was also informative of the factor structure, but with the huge impediment that it does not construct linear measures for even one of its many factorial dimensions. Further, it requires the analyst to determine which factor reflects the predominant measurement system, and which the multidimensionality. Factor rotation or other factor methods may clarify this, but they can also confuse the factor structure further (Ferguson, 1941).

In this study, Rasch analysis followed by factor analysis of residuals was always more effective at both constructing measures and identifying multidimensionality than direct factor analysis of the original response-level data.

References

Adams, R. J., and Wright, B. D. (1994) When does misfit make a difference? In M. Wilson (Ed.) *Objective Measurement: Theory into Practice, Vol. 2,* 244-270. Norwood, NJ: Ablex Publishing Corporation.

Andrich, D. A. (1991) Local independence, latent class analysis and Rasch. *Rasch Measurement Transactions, 5, 3,* 160-161.

Ferguson, G. A. (1941) The factorial interpretation of test difficulty. *Psychometrika 6, 5,* 323-329

Glas, C. A. W., and Verhelst, N. D. (1995) Testing the Rasch model. In G. H. Fischer and I. W. Molenaar (Eds.) *Rasch Models: Foundations, Developments and Recent Applications*, 69-96. New York: Springer Verlag.

Harman, H. H. (1960) *Modern Factor Analysis.* Chicago: University of Chicago Press.

Lazarsfeld, P. F. (1958) Latent structure analysis. In S. Koch (Ed.) *Psychology: A study of science. Vol. III,* 476-543. New York: McGraw-Hill.

Linacre, J. M. (1992) Prioritizing misfit indicators. *Rasch Measurement Transactions 9, 2,* 422-423.

Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V., and Hamilton, B. D. (1994) The structure and stability of the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation 75, 2,* 127-132.

McDonald, R. P. (1985) *Factor Analysis and Related Methods.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Smith, R. M., and Miao, C. Y. (1994) Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.) *Objective Measurement: Theory into Practice. Vol. 2,* 316-327. Norwood, NJ: Ablex Publishing Corporation.

Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological Measurement, 48*, 657-667.

Smith, R. M. (1996) A Comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling 3, 1,* 25-40.

Thurstone, L. L. (1932) *The Theory of Multiple Factors.* Ann Arbor, Michigan: Edwards Brothers.

Wilson, M., and Wang, W.-C. (1995) Complex composites: issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19, 1*, 51-71.

Wright, B. D. (1995) Scores, reliabilities and assumptions. *Rasch Measurement Transactions, 5, 3*, 157-158.

Wright, B. D. (1996) Comparing Rasch measurement and factor analysis. *Structural Equation Modeling, 3, 1,* 3-24.

Wright, B. D., and Linacre, J. M. (1997) *BIGSTEPS Rasch analysis 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.

Acknowledgement

Suggestions made by Mark Wilson and two anonymous referees have improved this paper.

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.

 ISSN 1090-655X.