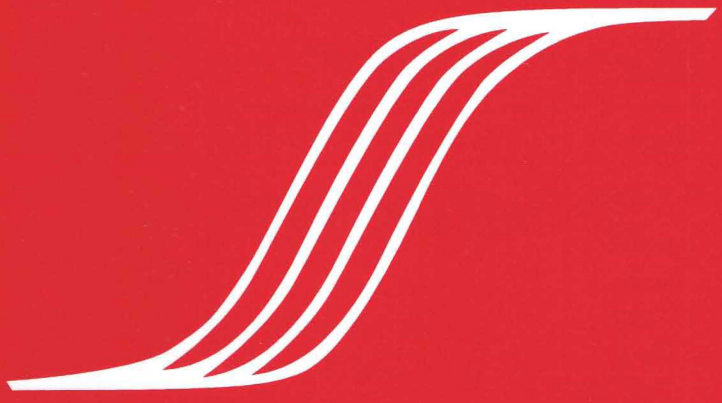
Volume 3, Number 1, 1999 ISSN 1090-655X

Journal of

Outcome Measurement®

Dedicated to Health, Education, and Social Science



REHABILITATION
FOUNDATION
INC. EST. 1993
Research & Education

EDITOR

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

ASSOCIATE EDITORS

Benjamin D. Wright . . . . . University of Chicago
Richard F. Harvey . . RMC/Marianjoy Rehabilitation Hospital & Clinics
Carl V. Granger . . . . . State University of Buffalo (SUNY)

HEALTH SCIENCES EDITORIAL BOARD

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

EDUCATIONAL/PSYCHOLOGICAL EDITORIAL BOARD

David Andrich . . . . . .Murdoch University
Trevor Bond . . . . . .James Cook University
Ayres D'Costa . . . . . Ohio State University
Barbara Dodd . . . . . University of Texas, Austin
George Engelhard, Jr. . . . . Emory University
Tom Haladyna . . . . . Arizona State University West
Robert Hess . . . . . Arizona State University West
William Koch . . . . . University of Texas, Austin
Joanne Lenke . . . . . Psychological Corporation
J. Michael Linacre . . . . . .MESA Press
Geofferey Masters . . . . . Australian Council on Educational Research
Carol Myford . . . . . Educational Testing Service
Nambury Raju . . . . . Illinois Institute of Technology
Randall E. Schumacker . . . . . .University of North Texas
Mark Wilson . . . . . .University of California, Berkeley

JOURNAL OF OUTCOME MEASUREMENT®

Volume 3, Number 1 1999

Articles

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

JOURNAL OF OUTCOME MEASUREMENT®, *3*(1), 1–20
Copyright© 1999, Rehabilitation Foundation, Inc.

A Comparison of Three Polytomous Item Response Theory Models in the Context of Testlet Scoring

Karon F. Cook
Baylor College of Medicine/Veterans Affairs

Barbara G. Dodd
Steven J. Fitzpatrick
University of Texas at Austin

An alternative to dichotomous scoring of multiple items anchored to a common stem is scoring these items as a single polytomous item (testlet scoring). This study systematically compared the partial credit model (PCM), the generalized partial credit model (GPCM), and the graded response model (GRM) in the context of testlet scoring. Data sets included a sample from the fall 1994 administration of the SAT I (N=2,548) and a simulated data set.

Theta estimation, information, and model fit were analyzed. Correlations among theta estimates ranged from 0.9748 to 0.9921. The relationship among the information functions of the PCM, GPCM and the GRM reflected the discrimination parameter estimates for the latter two models. Suggestions are made with regard to model selection.

Requests for reprints should be sent to Karon F. Cook, Houston Rehabilitation R&D Field Program, VA Medical Center (153), 2002 Holcombe Blvd., Houston, Texas 77030, e-mail: karonc@bcm.tmc.edu
The preparation of this manuscript was funded partially by a post-doctoral fellowship from Veterans Affairs Health Services Research & Development.

An assumption of most item response theory models (IRT) is unidimensionality, and an important consequence of this assumption is local item independence. Local independence exists when, for any given level of the trait being measured, responses to items are statistically independent (Hambleton and Swamiinathan, 1985). The weaker form of local independence assumes that, for any given level of the trait being measured, responses to different items are uncorrelated.

Even if the content of an IRT-calibrated test meets the assumption of unidimensionality, however, it is possible to violate local item independence through the use of an inappropriate scoring rubric. Such a violation is apparent in tests which include sets of items that share a common anchor such as a reading passage, a graph, or a portion of computer code. When such items are scored individually as dichotomous items (right/wrong), the items within a subset may exhibit local dependence, and, therefore, be inconsistent with the assumptions of unidimensional IRT models. When there is a small number of items anchored to a passage, the unidimensional models may be robust to such violations. The trend in ability testing, however, is toward subsets of as many as twelve or thirteen items anchored to a common stem. In such situations, local dependence among items may be problematic.

Local dependence among test items has substantial consequences in IRT applications. Research has shown that local item dependence can result in inaccurate estimation of item parameters, information functions, and reliability (Sireci, Thissen, and Wainer, 1991; Thissen, Steinberg and Mooney, 1989; Wilson, 1988; Yen, 1984; Yen, 1993). An alternative to dichotomous scoring of items sharing a common anchor is to consider each subset of items as a single fungible unit or "testlet." The testlet score is the total number of individual items within the subset that an examinee correctly answers. Wainer and Kiely (1987) first proposed an expanded use of testlets, noting that one advantage of such an approach is that it shifts the requirement of local independence from the item level to the testlet level. This is an expedient move since, as Rosenbaum (1988) has shown, local independence may exist at the testlet level even if it is violated at the item level.

A number of polytomous models have been applied in testlet scoring, and there is some indication that different results can be obtained depending upon which polytomous model is applied (Wilson, 1988). The purpose of this research was to compare three polytomous IRT models in

the context of testlet scoring: the graded response model (GRM) (Samejima, 1969), the partial credit model (PCM) (Masters, 1982), and the generalized partial credit model (GPCM) (Muraki, 1992).

Models

Graded response model

In 1969 Samejima proposed the GRM, an extension of the dichotomous two-parameter logistic (2-PL) model, to the multiple category case. In the GRM, responses to item i are classified into $m_i + 1$ categories, where m represents the highest possible score on item i. The set of possible scores on item i, is defined as $(0, 1, \ldots, m_i)$. The response categories are ordered such that higher category scores represent more of the trait being measured than do lower scores.

Thissen and Steinberg (1986), who proposed a taxonomy of IRT models, classified the GRM as a "difference model." In a difference model the probability of an examinee of given trait level scoring in a particular response category is not obtained directly. Samejima (1969) defined a two stage approach to finding this probability. First, the probability of an examinee with a given trait level scoring in a given category or higher is defined as

$$P^*_{ix}(\theta) = \frac{\exp[Da_i(\theta - b_{ix})]}{1 + \exp[Da_i(\theta - b_{ix})]}, \quad (1)$$

where

b_{ix} = the category boundary for score x on item i,

a_i = the discrimination parameter for item i, and

D = scaling constant (1.7).

The category boundary parameter (b_{ix}) is the difficulty parameter associated with category score x for item i. It may be thought of as the difficulty of getting this category score or one higher. For item i, there are $m + 1$ possible category responses and m category boundaries. For the homogenous case of the GRM, the discrimination parameters are assumed to be equal across all categories within an item. Equation 1 is used to determine $P^*_{ix}(\theta)$ (category characteristic function) for all category responses except for 0 or for $m + 1$. $P^*_{ix}(\theta)$ for extreme category scores are defined as follows

$$P^*_{io}(\theta) = 1 \quad (2)$$

and,

$$P^*_{i(m_i+1)}(\theta) = 0. \quad (3)$$

Equation 2 defines the probability of scoring in category 0 or higher as unity. Equation 3 defines as zero the probability of scoring higher than the highest category score.

The second step in determining the probability that an examinee of given trait level will score in a particular category requires the subtraction of adjacent category characteristic functions. Specifically, the probability that an examinee with a given trait level will score in a particular category is defined as

$$P_{i_x}(\theta) = P_{i_x}^*(\theta) - P_{i_{x+1}}^*(\theta). \quad (4)$$

Samejima defined the item information function for the GRM with the following equation:

$$I_i(\theta) = \sum_{x=0}^{m_i} \frac{[P'_{ix}(\theta)]^2}{P_{ix}(\theta)} - \sum_{x=0}^{m_i} P''_{ix}(\theta), \quad (5)$$

where

$I_i(\theta)$ = information for, item i, for examinees with trait level equal to θ,

$P_{ix}(\theta)$ = the probability of examinees of a given trait level, (θ), responding in category x,

P'_{ix} = the first derivative of $P_{ix}(\theta)$,

P''_{ix} = the second derivative of $P_{ix}(\theta)$.

As Samejima showed, the second term in Equation 6 is equal to zero and, therefore, drops out of the equation. The test information function is equal to the sum of the item information functions.

Partial credit model

In 1982 Masters introduced the partial credit model (PCM). As is the case with the GRM, the PCM is useful for items with more than two response categories. Like the GRM, the PCM assumes ordered category responses. Unlike the GRM, however, the PCM is what Thissen and Steinberg (1986) have classified as a "divide by total model." In divide by total models, the denominator is equal to the sum of all possible nu-

merators, and the probability of an examinee of given theta level scoring in a particular category is obtained directly. Another distinction between the PCM and the GRM is that the former is a member of the Rasch family and does not include a discrimination parameter in the model.

Masters considered the category responses associated with a given polytomous item as a series of successive "steps." An examinee either succeeds or fails each step within an item. An individual's category score is the sum of his or her step scores, i.e., the number of steps passed. Masters defined the probability of a given category score as

$$P_{ix}(\theta) = \frac{\exp[\sum_{k=0}^{x} (\theta - b_{ik})]}{\sum_{h=0}^{m_i} \exp[\sum_{k=0}^{h} (\theta - b_{ik})]} \quad (x = 0, \ldots m_i), \tag{6}$$

where

b_{ix} = the difficulty of the step associated with the category score, x, for item i, and,

m_i = the highest possible score on item i.

Though the response categories must be ordered when using the PCM, the step difficulties do not have to be ordered; i. e., reversals are allowed.

Generalized partial credit model

Muraki (1992) has proposed the generalized partial credit model (GPCM), an extension of the PCM. In contrast to the PCM, in the GPCM the discrimination parameter is allowed to vary across items. In the GPCM, the probability of a particular category score, x, given theta is defined as

$$P_{ix}(\theta) = \frac{\exp[\sum_{k=0}^{x} a_i(\theta - b_{ik})]}{\sum_{h=0}^{m_i} \exp[\sum_{k=0}^{h} a_i(\theta - b_{ik})]}, \tag{7}$$

where

a_i = the discrimination parameter for item i, and

b_{ik} = the difficulty of the step associated with category k, $(k = 1, \ldots, m_i)$.

As is the case with the PCM, reversals are allowed in the GPCM.

Method

Data Sets

Two data sets were employed in this study, one archival and one simulated. The archival data set was obtained from a fall, 1994 administration of the College Board's Scholastic Assessment Test I (SAT I) and was composed of the responses of 2,548 examinees to the verbal section of this test. The SAT I Verbal section included 19 analogies, 19 sentence completions and 40 critical reading questions. The critical reading questions were grouped into testlets, with between 5 and 13 items per testlet. The SAT I data set was included to allow comparisons among the PCM, GPCM, and GRM in an actual testing context.

To complement the SAT I data, a simulated data set (N=2000) was generated using a linear factor analytic approach suggested by Wherry,

Table 1

Input Factor Loadings for the Generation of the Simulated Data Set

Item #	Factor I	Factor II	Item #	Factor I	Factor II
1	0.42869	-0.35061	22	0.12980	0.05308
2	0.44546	-0.30539	23	0.23553	0.16661
3	0.47899	-0.29618	24	0.41396	-0.08526
4	0.42941	0.11230	25	0.50779	-0.17141
5	0.44974	0.13179	26	0.52690	-0.03497
6	0.38194	0.05426	27	0.44592	-0.05639
7	0.57085	0.20179	28	0.32087	-0.05456
8	0.46524	0.09206	29	0.43387	-0.01801
9	0.32207	0.02620	30	0.55128	0.08400
10	0.36134	0.15202	31	0.45327	0.09282
11	0.28833	-0.05559	32	0.43707	0.06332
12	0.32170	-0.13629	33	0.28765	-0.03655
13	0.37487	-0.14023	34	0.43899	-0.10330
14	0.25547	-0.03412	35	0.42145	0.03005
15	0.28145	0.02876	36	0.31636	0.07735
16	0.52793	0.01819	37	0.32793	0.03327
17	0.45729	0.00884	38	0.35608	0.11903
18	0.38849	0.04263	Testlet 1	0.77583	-0.08570
19	0.55508	0.01466	Testlet 2	0.65200	-0.8226
20	0.32466	0.05076	Testlet 3	0.80761	0.1349
21	0.31256	0.17257	Testlet 4	0.83078	0.0948

Naylor, Wherry, and Fallis (1965). The advantage of generating data using this approach over generating data to fit the models in the study is that the factor analytic approach does not prejudice the data toward one model or another. Data that are model-neutral allow for the clearest comparison among models. The generation program required the input of: (1) a factor loading matrix for the test items and (2) z-score cutting points. Since the purpose of this research was to make comparisons which would generalize to practical testing situations, an effort was made to simulate data that were similar to real data. To that end, the factor loadings for the simulation program were obtained from a principal axis factor analysis of the 38 dichotomous items and the 4 testlets of the SAT I Verbal data set (Table 1). The z-score cutting points for the items were chosen so that the frequencies of possible item scores approximated that of the SAT I Verbal data. Since the theta scale is similar in range to the z-score scale, the cumulative proportions of item scores from the SAT I Verbal data were converted to the z-scores associated with corresponding proportions under the normal curve.

Parameter Estimation

The PARSCALE (Muraki and Bock, 1993) software package was used to estimate the parameters of both the SAT I Verbal data and the simulated data sets. Parameter estimates were obtained using the GRM, GPCM, and PCM. PARSCALE employs a marginal maximum likelihood EM algorithm for item parameter estimation (Muraki, 1992). This algorithm consists of two steps: (1) provisional expected frequency and sample size are calculated, and (2) the marginal maximum likelihood is estimated. The steps are repeated until item parameter estimates stabilize. Finally, maximum likelihood is used to estimate person parameters.

Analyses

Comparison of Theta Estimates

For the SAT I Verbal data set, the Pearson product-moment (PPM) correlations among theta estimates from the GRM, GPCM, and PCM were calculated. For the simulated data set, the PPM correlations between each of the model's theta estimates and simulees' z-scores on the first factor were computed. Since the simulated data were generated to have a dominant first factor, use of the z-score for the first factor as a representation of theta level was deemed reasonable. In addition to PPM correlations,

the root mean squared errors (RMSEs) for each model's estimates were calculated for the simulated data. As with the PPM correlations, the z-score for the first factor represented a simulee's theta level. The difference between the model's estimate of theta and the z-score for the simulee on the first factor was defined as the amount of error for each simulee. These errors were squared and averaged across simulees. The square root of this quantity served as the RMSE.

Information

The third set of analyses was a comparison of the information functions from each model across the theta scale on each data set. Calculation of test information was accomplished using the IRTINFO program (Fitzpatrick, S.J., Choi, Chen, Hou, and Dodd, 1994). Before the information functions were compared, it was necessary to equate them (Fitzpatrick and Dodd, 1997). True score equating based on test characteristic curves (TCCs) is one of the most widely used IRT equating techniques (Stocking and Lord, 1983; Baker and Al-Karni, 1991; Baker, 1992; Baker, 1993). Because the item parameters of different models are not always defined in the same manner, however, this technique is not appropriate when different IRT models have been used to calibrate the same test. The GRM (Samejima, 1969), for example, requires the *b* parameters to be ordered, while the PCM (Masters, 1982) and the GPCM do not. Therefore we employed an equating approach developed by Fitzpatrick and Dodd (1997) for correcting the information function for the parameter scale transformation. The procedure involves the use of an unspecified monotonic transformation of the theta scale of one model to bring its TCC into correspondence with the TCC from the base model. An equation that models the transformation is then obtained and used to modify the information function of the model to be equated to place it on the same scale as the base model. The PCM served as the base model for the current study. For each data set, once the information functions were equated, information was compared across models.

Model Fit Analyses

In addition to the analyses described above, the three models were compared with regard to model fit. Hambleton and Swaminathan (1985) have suggested residual analysis as a way of assessing model fit. In this approach, the empirical probabilities of scoring in a given category are

Table 2
Item Parameter Estimates for Testlets 1-4 (T1-T4) of the SAT I Verbal Data Set

PCM	a	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10	b11	b12	b13
T1		-3.262	-2.680	-2.036	-1.801	-1.511	-1.364	-1.123	-0.712	-0.606	-0.029	0.326	0.858	2.223
T2		-2.155	-1.741	-0.931	-0.536	0.378								
T3		-3.102	-1.916	-1.334	-0.951	-0.343	-0.070	0.319	0.621	0.784	1.572	2.236		
T4		-2.541	-1.684	-1.369	-0.701	-0.436	-0.219	0.288	0.472	0.499	1.057	1.642		
GPCM														
T1	0.825	-3.441	-2.797	-2.070	-1.834	-1.533	-1.406	-1.168	-0.731	-0.671	-0.048	0.297	0.853	2.343
T2	0.921	-2.182	-1.779	-0.951	-0.579	0.360								
T3	0.927	-3.178	-1.938	-1.346	-0.970	-0.349	-0.086	0.307	0.614	0.779	1.625	2.342		
T4	0.961	-2.550	-1.683	-1.379	-0.706	-0.450	-0.241	0.274	0.458	0.486	1.073	1.698		
GRM														
T1	2.231	-3.880	-3.140	-2.591	-2.193	-1.837	-1.511	-1.172	-0.804	-0.440	0.023	0.521	1.130	2.064
T2	1.624	-2.884	-1.976	-1.112	-0.323	0.751								
T3	2.541	-2.805	-1.985	-1.428	-0.962	-0.501	-0.103	0.291	0.679	1.086	1.647	2.319		
T4	2.722	-2.431	-1.801	-1.347	-0.896	-0.529	-0.185	0.175	0.491	0.814	1.270	1.865		

Table 3
Item Parameter Estimates for Testlets 1-4 (T1-T4) of the Simulated Data Set

PCM	a	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10	b11	b12	b13
T1		-2.629	-1.820	-1.498	-0.864	-0.832	-0.403	0.134	0.208	1.075	0.691	1.550	2.016	2.521
T2		-1.998	-0.730	0.073	0.841	1.872								
T3		-2.538	-1.546	-1.347	-0.798	-0.421	-0.014	0.407	0.792	1.230	1.967	2.578		
T4		-2.390	-1.694	-1.183	-0.892	-0.554	-0.291	0.268	0.459	0.879	1.251	2.537		
GPCM														
T1	0.5651	-3.597	-2.338	-1.924	-0.953	-1.040	-0.422	0.389	0.383	1.780	0.961	2.332	2.993	3.685
T2	0.7188	-3.988	-2.289	-1.238	-0.239	1.122								
T3	0.8749	-2.953	-1.772	-1.502	-0.833	-0.363	0.142	0.665	1.148	1.693	2.580	3.317		
T4	0.8773	-2.782	-1.943	-1.320	-0.951	-0.529	-0.193	0.482	0.739	1.261	1.731	3.243		
GRM														
T1	1.8363	-3.442	-2.545	-1.909	-1.308	-0.840	-0.325	0.200	0.664	1.220	1.617	2.230	2.909	3.729
T2	1.2562	-4.437	-2.639	-0.511	-0.076	1.404								
T3	2.2616	-2.870	-2.023	-1.453	-0.879	-0.356	0.153	0.671	1.200	1.767	2.464	3.270		
T4	2.3728	-2.753	-2.016	-1.459	-0.990	-0.540	-0.094	0.394	0.842	1.340	1.908			

calculated for different ranges of theta. Consider the example of a five category polytomous item. Suppose for examinees whose estimated theta falls between -1.0 and -0.5, 18% scored '0,' 22% scored '1,' 40% scored '2,' 15% scored '3,' 5% scored '4,' and no examinees scored '5.' The empirical probabilities for scores of '0' through '5' would be 0.18, 0.22, 0.40, 0.15, and 0.00, respectively for the theta range -1.0 to -0.5. The empirical probabilities are calculated in the same manner for each range of theta. The empirical and theoretical probabilities (category characteristic functions) are then plotted against theta, where theta is equal to the midpoint of the theta range for which the empirical probabilities are defined. The difference between the empirical and the theoretical probabilities is defined as the residual.

For the present study, empirical and theoretical probabilities were calculated for all testlets, all models and both data sets. The increment of theta used for calculating empirical probabilities was 0.5. In order to ensure stable probability estimates, empirical probabilities were not calculated for theta ranges in which less than 30 examinees fell. Because results were similar for each of the four testlets, only the probabilities for the five item Testlet 2 were selected to be presented graphically. This testlet was chosen because it had the fewest items and the plots of its theoretical and empirical probabilities, therefore, were the easiest to read.

Results

Item Parameter Estimates

The item parameter estimates from the testlets of the PCM, GPCM, and GRM calibrations are reported for the SAT I Verbal data and the simulated data in Tables 2 and 3, respectively. Item parameter estimates for the dichotomous items are available from the corresponding author upon request.

Table 4

Pearson Product-Moment Correlations among Theta Estimates for the SAT I Verbal Data Set

	PCM	GPCM	GRM
PCM	1.0000	0.9921	0.9919
GPCM		1.0000	0.9877
GRM			1.0000

Table 5

Pearson Product-Moment Correlations among Theta Estimates and First Factor z-Scores for Simulated Data Set

	PCM	GPCM	GRM	1st Factor z-Score
PCM	1.0000	0.9748	0.9883	0.9454
GPCM		1.0000	0.9813	0.9349
GRM			1.0000	0.9462
1st Factor z-Score				1.0000

Comparison of Theta Estimates

Table 4 reports the PPM correlations among the PCM, GPCM, and GRM theta estimates for the SAT I Verbal data set. Table 5 reports this information for the simulated data set. As the tables indicate, the correlations among the models were quite high. For the SAT I Verbal data, the correlations ranged from 0.9877 to 0.9921. For the simulated data, the correlations ranged from 0.9748 to 0.9883. The PPM correlations between the theta estimates of the various IRT models and the first factor z-scores of the simulated data also were quite strong, though not as strong as those found among theta estimates. These correlations ranged from 0.9349 to 0.9462. The RMSEs for the simulated data set were quite similar across models. The obtained values were 0.3157, 0.3301, and 0.3317 for the GRM, PCM, and GPCM, respectively.

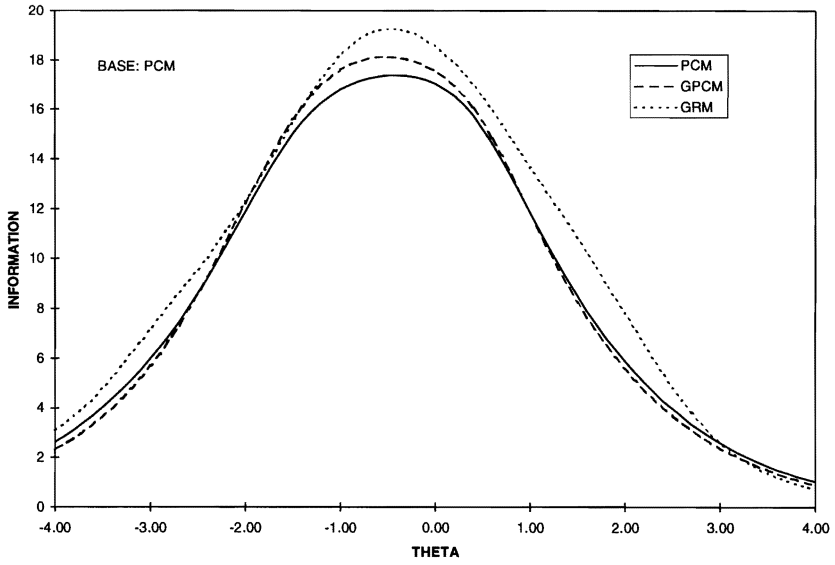


Figure 1. Test Information Functions for the SAT I Verbal Data Set.

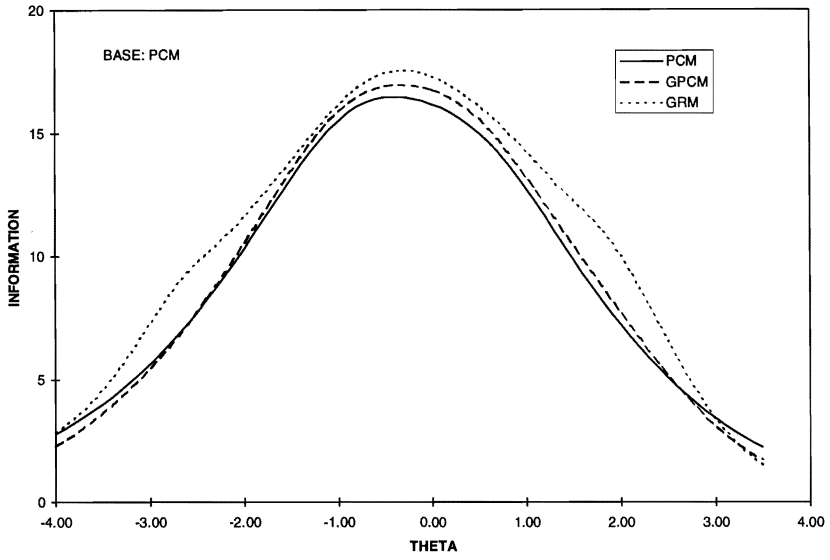


Figure 2. Test Information Functions for the Simulated Data Set.

Information Analyses

Figure 1 presents the test information functions for the calibrations of the SAT I Verbal data, after equating. Figure 2 presents this information for the calibrations of the simulated data set. The plots of the information functions for both data sets indicate that the GPCM yielded more information than the PCM at middle ranges of theta and that the GRM yielded more information than the PCM or GPCM across almost the entire range of the scale. The relationships among the various models' information functions is partly explained in comparing their item parameter estimates. Information functions are directly related to the slope of the item characteristic functions and the conditional variance at each level of theta. The steeper the slope and the smaller the conditional variance, the greater the information. The discrimination parameter is proportional to the slope of the item characteristic function. Therefore, tests whose items have greater discrimination parameter estimates tend to yield steeper information functions. Recall that the PCM does not include any item discrimination parameter (therefore, its value was set to 1.0 in the PARSCALE calibration). When the data do not fit this assumption of the PCM, the information functions will be artificially inflated or deflated depending upon whether the 1.0 estimate is an over- or an under-estimate of item discrimination. As indicated in Tables 2 and 3, in the GPCM, where the discrimination parameters were estimated, the values were below 1.0. This

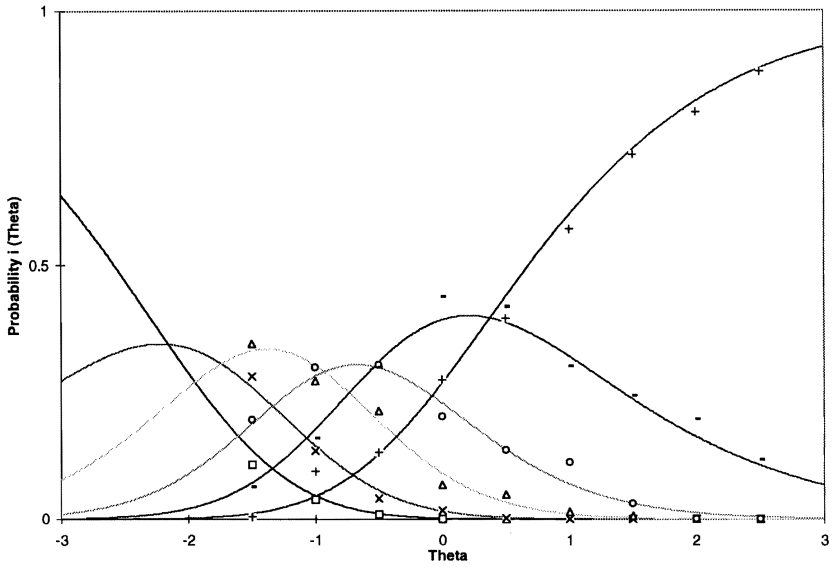


Figure 3. Empirical and Theoretical Probabilities for the PCM Calibration of Testlet 2 of the SAT I Verbal Data Set.

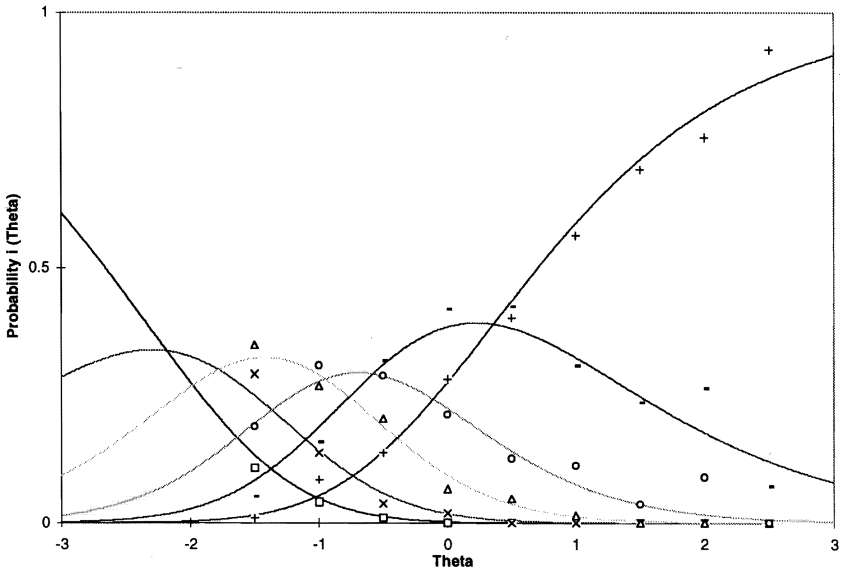


Figure 4. Empirical and Theoretical Probabilities for the GPCM Calibration of Testlet 2 of the SAT I Verbal Data Set.

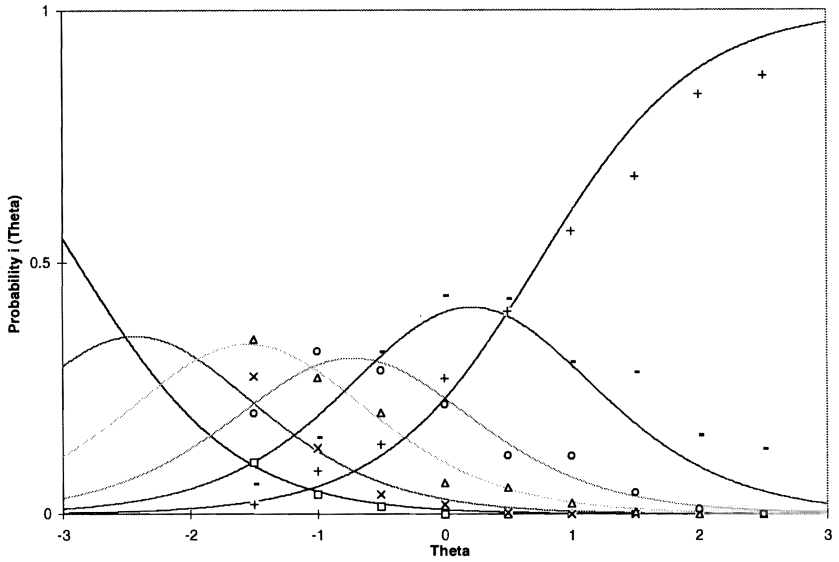


Figure 5. Empirical and Theoretical Probabilities for the GRM Calibration of Testlet 2 of the SAT I Verbal Data Set.

result suggests that the information functions of the PCM may be somewhat inflated because the model does not include nor estimate item discrimination parameters. The greater putative information obtained with the GRM can be explained in part by the relatively high value of the discrimination parameter estimates for the model. As reported in Tables 2 and 3, the discrimination parameter estimates for the GRM were substantially higher than those for the GPCM.

Model Fit Analyses

The empirical and theoretical probabilities for Testlet 2 are presented in Figures 3 through 8. As can be seen in Figures 3 and 4, the PCM and GPCM fit well the empirical probabilities of the SAT I Verbal data across all category scores, and the GRM fit well across most category scores (Figure 5). For a category score of '5,' however, the GRM theoretical probabilities were lower than the empirical probabilities for thetas less than or equal to 0.0. For theta values of 1.0 and above, the GRM theoretical probabilities were higher than the empirical probabilities for this category.

As can be seen in Figures 6 through 8, all three models fit the simulated data set well across categories, though the PCM fit a category score of '5' least well. In the PCM calibration, for a category score of '5,' the theoretical probabilities were higher than the empirical probabilities for

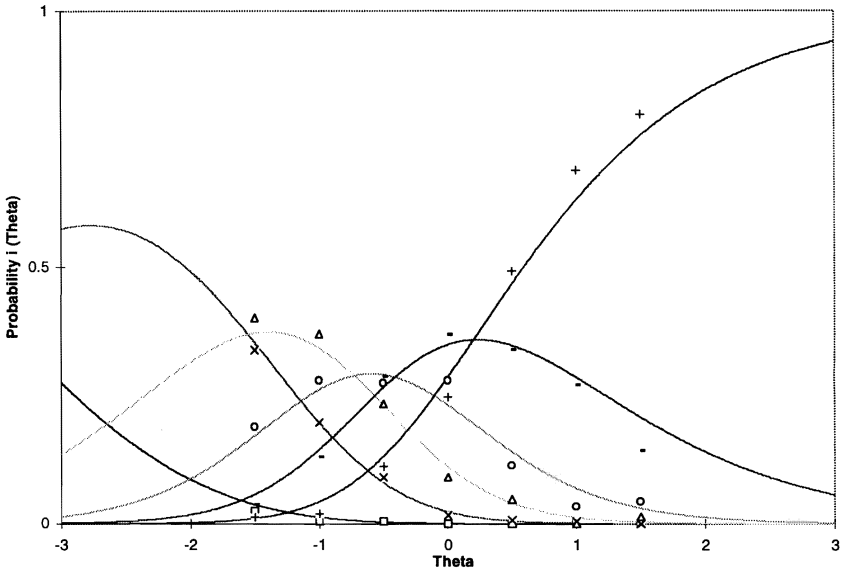


Figure 6. Empirical and Theoretical Probabilities for the PCM Calibration of Testlet 2 of the Simulated Data Set.

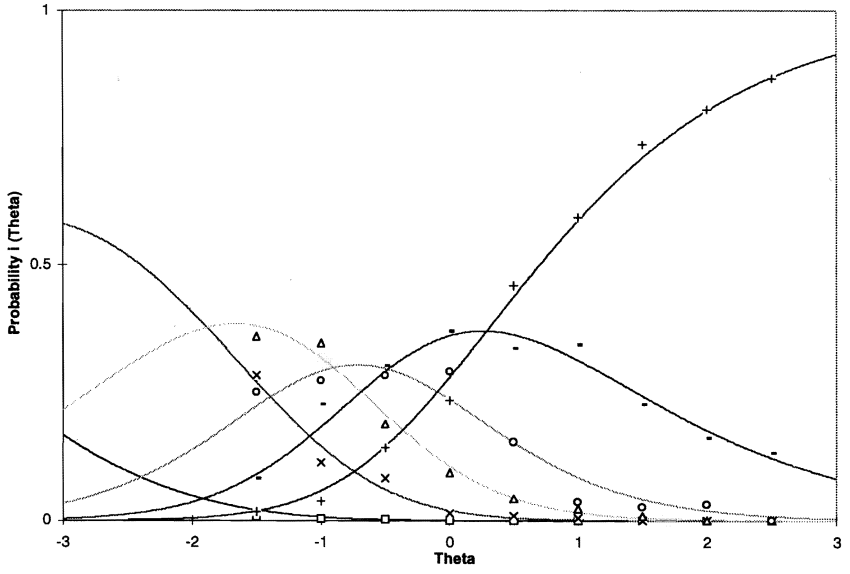


Figure 7. Empirical and Theoretical Probabilities for the GPCM Calibration of Testlet 2 of the Simulated Data Set.

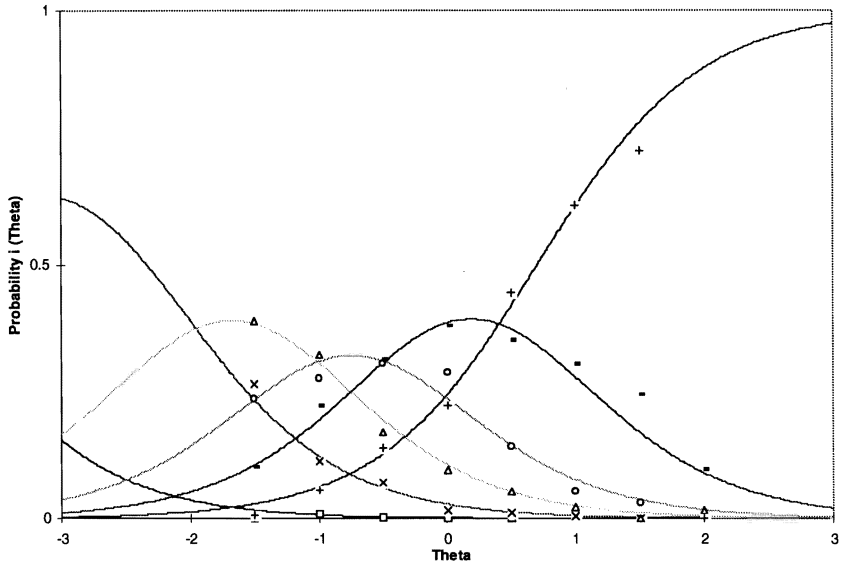


Figure 8. Empirical and Theoretical Probabilities for the GRM Calibration of Testlet 2 of the Simulated Data Set.

thetas less than 0.0. The theoretical probabilities were lower than the empirical probabilities for thetas of 1.0 and greater.

Discussion

The results of this study suggest that, with regard to the estimation of theta, the PCM, GPCM, and GRM all perform equally well in the context of testlet scoring. For both data sets, the three polytomous IRT models yielded theta estimates that were strongly associated with each other, and, for the simulated data set, strongly associated with their corresponding first factor z-scores. The RMSEs, calculated for the simulated data set, ranged from 0.3157 to 0.3317. The correlations for the simulated data set were lower and the RMSEs were higher than the results obtained by Dodd (1984) when she compared polytomous IRT models in the context of attitude scaling. However, the scale that Dodd employed had 30 polytomous items and, since test information is cumulative, this scale could be expected to perform as well as or better than the items used for the present study.

All three models compared in this study were found to exhibit good model to data fit across data sets as assessed by the plots of empirical and theoretical probabilities. The PCM performed as well or better than the GPCM and the GRM for the SAT I Verbal data set The GPCM and the GRM fit the simulated data set slightly better than did the PCM.

The similarity in the fit of the PCM with the other two models may be due, in part, to the similarity in the discrimination of the four testlets. As shown in Table 1, the first factor loadings obtained for the four testlets of the SAT I Verbal data were similar, ranging from 0.65 to 0.83. Since the testlets of the simulated data set were generated based on these factor loadings, the simulated testlets would also be expected to be similar in discrimination. The discrimination parameter estimates based on the GPCM and the GRM calibrations are somewhat similar in range, though this is much more the case for the SAT I Verbal data (GPCM: 0.83 to 0.96, GRM: 1.62 to 2.72) than for the simulated data (GPCM: 0.57 to 0.88, GRM: 1.26 to 2.37).

Based solely on the comparisons of theta estimation and model to data fit, there appears to be no clear advantage in selecting one model over the other. The one analysis which yielded notable differences among models was the comparison of information functions. The calculated information for the GRM was greater than either the PCM or GPCM across most of the range of theta. It is intriguing that this clear advantage in putative information did not express itself, however, in substantially stronger correlation coefficients or smaller RMSEs in comparison to the PCM and GPCM. The practical implications of the slightly greater information observed with the GRM should be explored.

All three models performed well in theta estimation. With regard to model selection, however, Andrich (1995) has argued that theoretical considerations, as well as empirical, should govern the choice of a model for a given psychometric application. He contends that the model should be consistent with the underlying response process. Andrich describes how common uses of what he calls "Thurstone models," including the GRM, are "internally inconsistent" because of a mismatch between the response process that occurs and the one that is modeled (e.g., joining assumption holds, trait level estimate is not explicitly separated from the location of response category thresholds, etc.). In the current context, based on the empirical findings, theoretical considerations, and parsimony, a case can be made for choosing the PCM over the GPCM or GRM.

Conclusions

The models examined in this study were compared using several indices and approaches, but no single set of analyses was sufficient to describe the relationships among the models. It is concluded, therefore, that model selection is most appropriately undertaken after considering a number of

factors. As Andrich has described, there should exist a consistency between the model and the response process (Andrich, 1995). Assessment of the degree to which data sets meet model assumptions is also an important consideration, especially since violation of these assumptions may affect putative information. Plots of empirical and theoretical probabilities are useful in assessing model fit. When there is sufficient model to data fit, information functions are helpful in comparing models. For simulated data sets, both RMSE and PPM correlations should be considered in evaluating the accuracy of models with regard to theta estimation.

Limitations and Suggestions for Future Research

The results of this study are limited by the choice of data. The SAT I is a well-constructed, high-stakes exam and the simulated data were modeled upon it. Comparison of polytomous IRT models in other contexts should be undertaken to ascertain the degree to which the results obtained in the present study are consistent across a variety of psychometric settings. The results of this study suggest a number of areas which warrant further exploration. These include the effect on local dependency of testlet length and of the ratio of number of testlet items to total number of items; the robustness of different models to local dependency, particularly with regard to item parameter estimation; the success of different polytomous models in estimating theta at different ranges of theta; and, the relationship between violations of model assumptions, information, and model fit.

References

Andrich, D. (1995). Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measurement, 19*(1):101-119.

Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17*, 239-251.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16*, 87-96.

Baker, F. B. and Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147-163.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more categories. *Psychometrika, 37*, 29-51.

Dodd, B. G. (1984). Attitude scaling: A comparison of the graded response and partial credit latent trait models (Doctoral dissertation, University of Texas at Austin, 1984). *Dissertation Abstracts International, 45*, 2074A.

Fitzpatrick, S. J., Choi, S. W., Chen, S., Hou, L. and Dodd, B. G. (1994). IRTINFO: A SAS macro to compute item and test information. *Applied Psychological Measurement, 18*, 390.

Fitzpatrick, S. J. and Dodd, B. G. (1997). *The effect on information of a transformation of the parameter scale.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston: Kluwer-Nijhoff Publishing.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E., and Bock, R. D. (1993). *The PARSCALE computer program* [Computer program]. Chicago, IL: Scientific Software International.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53*, 349-359.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No. 17.

Sireci, S. G., Thissen, D., and Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Thissen, D. and Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.

Thissen, D., Steinberg, L., and Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*, 247-260.

Wainer, H. and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wherry, R. J., Sr., Naylor, J. C., Wherry, R. J., Jr., and Fallis, R. F. (1965). Generating multiple samples of multivariate data with arbitrary parameters. *Psychometrika, 30*, 303-313.

Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement, 12*, 353-364.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

JOURNAL OF OUTCOME MEASUREMENT®, *3*(1), 21-34
Copyright© 1999, Rehabilitation Foundation, Inc.

The Development of a Practical and Reliable Assessment Measure for Atopic Dermatitis (ADAM)

Denise Charman
Victoria University of Technology

George Varigos
Royal Children's Hospital, Victoria, Australia

David J. de L. Horne
University of Melbourne

Frank Oberklaid
Centre for Community Paediatrics, Victoria, Australia

Previous measures of Atopic Dermatitis (AD) have not been adequate for research purposes. This paper describes a study conducted in dermatology clinics of the Royal Children's Hospital, Melbourne, Australia, to develop a reliable, valid and practical measure. A pool of items to describe both site and morphology of AD was generated from a literature survey and expert opinion. Selected items were incorporated into a measure with each item rated on a four point scale. The measure was piloted and revised to a simpler format and called the Atopic Dermatitis Assessment Measure (ADAM). Unidimensionality was established. Reliability was determined by comparing two doctors blind ratings on 51 patients (mean age=70 months). Agreement varied depending upon site and morphology with more agreement on "mild" AD than on "severe" AD. These results imply that operational definitions of the scales need to be defined more clearly. The measure satisfies the assumptions for a partial credit analysis.

Requests for reprints should be sent to Denise Charman, Victoria University of Technology, P. O. Box 14428, MCMC, Melbourne, 8001, Victoria, Australia. e-mail: denisecharman@vut.edu.au

Introduction

In clinical medicine, the understanding of grades of severity of disease has generally been implicit and acquired during an apprenticeship style of clinical training. Yet, in medical research, it is necessary to have ratings of grades of actual severity which are explicit, well-defined and on which medical raters agree.

One example of a disease in which severity has been poorly measured and grades of severity poorly defined is Atopic Dermitis (AD), often referred to as "atopic eczema". AD itself has been variously described. It is essentially a skin disease which is "a specific dermatitis in the abnormally reacting skin of the atopic, resulting in itch with sequelae as well as inflammation" (Rajka, 1986).

There have been a number of attempts to develop both a measure for overall AD severity and an index to define grades of severity. One of the most recent attempts has been described in the consensus papers from the European Task Force on Atopic Dermatitis and is provided on the web-site, *http://adserver.sante.univ-nantes.fr/Scorad_Course/CD-ROM.html.* These papers and the web-site describe a severity measure and index for grading AD severity, referred to as SCORAD (European Task Force on Atopic Dermatitis, 1993; Kunz, et al., 1997). This work was being developed coincidentally with that described in this paper.

Scoring systems have generally been vague about the AD body sites and AD morphology and/or scored extent separately from morphology. Both are important to clinical assessment and may well influence how patients or parents (on behalf of their children) adhere to treatment (Koblenzer, 1989; Gil, et al., 1988; Noren and Melin, 1989; Ewing, et al., 1991). In one study, severe AD was defined behaviourally, as AD requiring outpatient visits every three months (Devlin and David, 1992). Where AD severity has been rated, reliability and validity of the ratings generally have not been established.

One of the more precise scoring systems (Atopic Dermatitis Area and Severity Index, ADASI) was that by Bahmer, Schubert and Schafer (1991) using a detailed body chart. The total severity score was based on the sum of grids (superimposed over a diagram of a body) using a colour code for severity. They provided no information about standardisation of this grid system, nor information about reliability, especially inter-rater reliability of the coding system. Though, later, ADASI (Atopic Dermati-

tis Area and Severity Index) was shown that it could be used to detect skin changes in AD severity (Bahner and Schafer, 1992).

The reliability of SCORAD was estimated by determining level of agreement between ten trained clinicians evaluating ten slides on five morphological features. The results showed that reliability varied across the different scales and some unreliable scales were maintained in the final SCORAD version. For example, one of the morphological features, mild edema/papulation had agreement "corresponding to 54% probability graded identically by 2 clinicians randomly chosen" (p.25). Moreover, the 5 morphological features were not entirely consistent with those set out by standard work by Hanifin (1989), though even this latter work was not without criticism.

SCORAD and the Hanifin index had each morphological feature rated on a scale of none, mild, moderate and severe. These features, however, were to be graded with no reference to body sites that were scored separately as AD extent. Also, the criteria for each of these gradations were not operationally defined in words; for example, the definition of "severe" (in contrast to "moderate") was not provided. Photographs were provided by the European Task Force in Atopic Dermatitis (1993). However, it was acknowledged that some features were difficult to grade from photos. Thus, inter-observer reliability and validity of the early index and the recent SCORAD system continue to remain questionable (Jemec and Wulf, 1997; Sprikkelman, et al., 1997).

Costa, Rilliet, Nicolet and Saurat (1989) developed a system and took steps to investigate the reliability of two summary scores, one of which was "simple" (8 morphology items and 10 sites) and the other more "complex" (6 morphology items and 20 sites). They assessed a small number of patients (n=14), over an unspecified number of visits for each individual, for a total of 100 visits. Items were included for pruritus, loss of sleep and for global "evolution" indicating any change from remission to worsening. Results for the simple system showed an approximately normal distribution, whereas the distribution for the complex one was positively skewed. That is, the simple scale was more statistically satisfying. They did not say which distribution more accurately reflected clinical presentations. The simple and complex scales were highly correlated (r=.90). Using the Wilcoxon signed rank test, however, showed that "the more elaborate method was less reproducible when two physicians' results were compared" (p. 45). Thus, Costa, et al. (1989) did acknowledge

the need for a reliable measure, but also demonstrated that it was difficult to obtain one.

Unless a well designed severity measure can be shown to be reliable and valid, outcome studies will be compromised. There will continue to be a paucity of information about diagnostic and descriptive variation as a contaminator in empirical research (and clinical judgement). Moreover, it will not be possible to develop an index for grades of severity nor reliably document the pattern of AD, according to age and sex of patients. Clinical evaluations of the disease will not be verifiable.

Thus, the aim of the present study is to develop a reliable and valid measure of the AD severity that can be used both in research and by busy medical practitioners. The severity measure can also be an index capable of detecting patterns of distribution and morphology, in each phase of the disease and at each developmental stage.

The Development of the Measure

The best solution to the search for a brief clinical measure was a paper and pencil measure. Physical and laboratory findings sometimes yield variable results dependent upon operator skill and type of measure (Agner et al., 1990) and are no more reliable (Sackett et al., 1991). It was decided that items for the measure were to be drawn from reports by dermatologists (expert opinion) and from a literature review, particularly from the work of Hanifin and Lobitz (1977) and Rajka (1989).

The morphological features chosen were scale/dryness(S), lichenification /thickness(L), erythema/redness(Er) and excoriations/scratch marks(Ex) summarized by the acronym, SLEE (coined by one of the authors, G.V.). Each morphological feature was to be rated for site and severity. Global items for the extent of each body area with AD and a severity estimate for each body area were also to be included. These requirements meant that the initial pool had 176 items, but this was reduced to 77 items to simplify the measure for use in clinics. Scaling methods were reviewed (Kline, 1986; Streiner and Norman, 1989) and the scale adopted was "none, mild, moderate and severe".

Thirty-eight items related to personal and family history of atopy and other aspects not descriptive of the actual clinical manifestation of AD were not included the analysis. This paper will focus on the 49 (7 x 7) site and morphology items that were located on the measure in a sequence that would be consistent with a clinical examination of a patient. The sequence was,

face, arms, hands, legs, feet and front and back. Each site was coded from 0 (none) to 3 (severe) for scale, lichenification and erythema. Excoriations were counted and coded as less than 5 (coded as 1) to more than 20 (coded as 3). Flexures (folds) for head and neck and limbs were coded separately as either 0 (absent) or 1 (present). Similarly the areas of scalp and napkin area (for infants) were coded as absent or present.

This paper and pencil measure was piloted to determine how "manageable" it was for the clinician and for ease of completion. The pilot study sample consisted of 15 children (age 3 months to 15 years) who attended a public specialist dermatology clinic. The parents of the patients gave informed consent for the treating doctor to assess their children. Three doctors administered the measure on 15 children. Five children were assessed on more than one occasion giving a total of 20 assessments. One doctor completed four, one completed nine and one completed seven assessments. The treating doctor was advised to complete the measure, to describe the AD "as it is now".

Visual appraisal of the completed measures showed that missing data could represent true negatives or may be due to the format. Each measure also took a long time to complete and codes were inconveniently placed. Consequently, the number of items was reduced to 29 and the format revised. This new measure was given the name ADAM, an acronym for Atopic Dermatitis Assessment Measure (see Figure 1). Operational definitions of "mild", "moderate" and "severe" were generated at a meeting of staff of the Dermatology Unit. As this was a time consuming occupation for preoccupied clinicians, discrepancies in criteria and some variations in the content of the photographs were accepted. Definitions of pruritus were adopted from Rajka and Langeland (1989). All definitions were put together in a manual which is available from the first author. Photographs were located by dermatologists to enhance their definitions. In order to evaluate this measure, a series of studies was undertaken.

Study 1

Study 1 was to determine if the ADAM measure could provide an adequate summary of doctors' ratings of AD severity. There appear to be two conventional ways of developing a summary scale. Firstly, all rated items could be summed, however, this common practice is not considered appropriate (Zhu, 1996). Secondly, each item could be analysed and a factor analysis conducted to assess uni-dimensionality.

Date: ..... **Name:**.....

Completed by: ..... **UR:** .....

DOB: .....

Sex. M=1 F=2

1. PRURITUS:

Please circle the appropriate word

None=0 Mild=1 Moderate=2 Severe=3

2. SITE & MORPHOLOGY:

Please place the appropriate number in each box

	0=None, 1=mild 2=moderate, 3=severe			0= <5, 1=5-20 2= >20, 3=fissured
	Scale Dryness	**Licheni- fication**	**Erythema**	**Excori- ations**
FACE	[___]	[___]	[___]	[___]
ARMS	[___]	[___]	[___]	[___]
HANDS	[___]	[___]	[___]	[___]
LEGS	[___]	[___]	[___]	[___]
FEET	[___]	[___]	[___]	[___]
TRUNK	[___]	[___]	[___]	[___]

Please circle the appropriate word, for these sites:

SCALP	Present	Absent	
NAPKIN AREA	Present	Absent	
FLEXURES:	HEAD & NECK	Present	Absent
	LIMBS	Present	Absent

3. GLOBAL RATING OF SEVERITY:

Please circle the appropriate word

None=0 Mild=1 Moderate=2 Severe=3

Figure 1. ADAM: Atopic Dermatitis Assessment Measure (©Denise Charman, George Varigos 1997)

Method

Participants: Children (N=171) with active AD who were consecutive patients to dermatology clinics at the RCH over a twelve month period were recruited. The children were new and old patients to the clinic (mean age=54 months, ranging from 4 to 193 months). There were 98 (57.3%) males (mean age=47 months) and 68 (39.8%) females (mean age=62 months). Five (2.9%) children did not have their sex recorded and 19 children did not have their age recorded. *The doctors*: There were three dermatologists and two dermatology trainees supervised by the Head of the Unit.

The ADAM measure: The ADAM measure comprises items scored on either a four point rating scale ("none", "mild", "moderate" or "severe") or a two point scale ("present" or "absent"). The doctors were advised when rating the site and morphology section, they could leave a blank rather than a zero when AD was absent.

Procedure: Participants were assessed in the dermatology clinics of the Royal Children's Hospital in Melbourne. Their treating doctor was instructed to rate the AD "as it is now".

Results

The scored ADAM measures were examined to determine if participants met the criteria for the diagnosis of AD. All children had some sites of the body with some morphological features of AD. Eleven (6.5%) children had no record of family history for eczema, asthma or hay fever. Sixty-six (38.4%) had a family history of asthma, 63 (36.6%) had a family history of hay fever, and 68 (39.5%) had a family history of eczema. Only 11 (6.5%) children had no family history of atopy. Personal histories showed 14 (25%) had asthma and 6 (11%) had hay fever. Five (10%) had other skin problems, and 14 (25%) had health problems, including allergies, diarrhea, eye cancer, and a heart murmur.

All item frequency distributions were positively skewed with high frequencies for zero responses. However, the pattern of zero frequencies did not reduce to SLEE, the order of the columns on the ADAM measure. Nor was the pattern of zero frequencies related to the order of body sites. Thus, no response sets were evident.

There were low frequencies for "severe" in the site and morphology section, with frequencies ranging from 1 (face/lichenification) to 24 (hands/

erythema). AD appeared more often on the arms and legs and less often on feet and hands. AD on the face appeared as moderate scale or erythema and less often as lichenified or excoriated. Arms and legs had lichenification and were moderately excoriated. The trunk had scale, but was less frequently excoriated or lichenified. The assessment of trunk required undressing the child and yet this site was as likely to be endorsed as face. These patterns concur with clinical impressions. Overall, morphological features were evident in decreasing order from erythema, scale, excoriations to lichenification.

On the summary global severity rating item, there were 74 (43%) with "mild" eczema, 82 (48%) with "moderate eczema and 12 (7%) with "severe" eczema. Of the five (3%) children admitted to hospital on the day of assessment, three had "severe" global ratings and two had "moderate" ratings. All the subjects were followed for six months from the day of assessment. During this time, 20 children were admitted for unspecified reasons. Three (15%) children had "severe" global ratings, 11 had "moderate" and 6 had "mild" ratings. There were 9 patients with "severe" global ratings who were not admitted. In short, there was no pattern evident between global rating and being admitted.

As a measure of severity, the disparate items on the ADAM measure need to constitute a uni-dimensional scale defined as "the existence of one latent trait underlying the data" (Hattie, 1985, p.157). The following hypothesis was proposed: It will be possible to identify one factor to represent the relationships among the items on the ADAM measure and the item test score correlations will be reasonably homogeneous (refer to Hambleton, Swaminathan and Rogers, 1991).

To analyze the data further, moderate and severe codes were collapsed into one and the ADAM items correlated. Missing data were recoded as zeros in accordance with instructions to assessing doctors. The correlation matrix was analyzed with Principal-Components analysis, using listwise deletion of cases. The plot of eigenvalues was examined to determined whether a dominant first factor was present (Hambleton, Swaminathan and Rogers, 1991). A "break" in the eigenvalues was easily be identified. The first eigenvalue was 7.20 (24.8% of the variance), then the eigenvalues reduced to 2.73 (9.4%), 2.29 (7.9%), 1.9 (6.7%) and so on. All factor loadings were .41 and above except for the face items where factor loadings ranged from .21 to .31. The items loading onto the second factor were the scale dryness (S) items. The items loading onto the third factor were the face items.

Discussion

The ADAM measure did provide the required specific information for a diagnosis of AD and detected differences in the type and severity of morphological features of AD detected on body sites. Interestingly, the global ratings of severity were not associated with hospital admission.

The general factor, expressed as the percentage of total variance (28%) was taken as evidence of uni-dimensionality (Hattie, 1985). A further and alternative index for uni-dimensionality which is the difference between the first and second eigenvalues divided by the difference between the second and third eigenvalues. This ratio is 4.46/.34 = 13.1. Thus, the ADAM measure can be considered to be uni-dimensional.

Study 2

Since total scores on rating scales such as the ADAM measure may be misleading and not to be used to estimate reliability (Zhu, 1996), this study of the reliability continued the examination of independent items to determine the level of inter-rater agreement. The ADAM measure requires visual assessment of AD without any physical or laboratory findings. Under these circumstances, the question needs to be asked, how well do doctors agree on grades of severity, that is, "none", "mild", "moderate" and "severe"?

Method

Participants: Fifty-one children who presented to dermatology clinics at RCH, who had active AD diagnosed by specialist dermatologists. Each child was assessed by two doctors. There were 31 (61%) male and 20 (39%) female children, aged between 5 and 161 months, M=70.3, SD=53.7, n=42, (age not recorded for 9 observations). *Doctors.* There were three dermatologists and two dermatology trainees supervised by the Head of the Unit.

Procedure: A doctor completed the measure to assess a child's AD, and then requested a second doctor, who was blind to the first assessment, to assess the child. The two assessments were done within a half hour period. The frequencies with which doctors provided data were 16 (15%), 32 (31%), 10 (10%), 25 (24%) and 19 (18%).

Analysis: Pooled kappas which were not weighted (Fliess, 1981) were computed with True Epistat Statistical Package (Gustafson, 1991). One-tailed test of significance was adopted since it was expected that

Table 1

Kappas for Sites and Morphological Items

Item		Pooled Kappas K	SE
Pruritus		.60*	.12
Face	Scale	.45*	.10
	Lichenification	.21***	.10
	Erythema	.34*	.10
	Excoriations	.51*	.11
Arms	Scale	.41*	.10
	Lichenification	.29*	.10
	Erythema	.34*	.10
	Excoriations	.57*	.10
Hands	Scale	.50*	.11
	Lichenification	.26*	.11
	Erythema	.45*	.11
	Excoriations	.52*	.11
Legs	Scale	.40*	.10
	Lichenification	.26	.09
	Erythema	.37*	.10
	Excoriations	.51*	.10
Feet	Scale	.47*	.14
	Lichenification	.33*	.11
	Erythema	.32*	.11
	Excoriations	.35*	.11
Trunk	Scale	.13	.14
	Lichenification	.30*	.10
	Erythema	.38*	.09
	Excoriations	.38*	.10
Scalp		.78**	.09
Napkin area		.56**	.13
Head/neck flexures		.39**	.17
Legs/arms flexures		.64**	.15

* $p<.01$

** $p<.05$ The upper and lower 95% confidence intervals for kappa did not include zero and can be considered as significantly different from zero

*** $p<.02$

both doctors would be scoring in the same direction. Ninety-five percent confidence intervals were also used. Currently, there is no agreed absolute value for significance of kappa beyond which it can be claimed that there is significant agreement between raters. However, the criterion kappa greater than .70 was chosen (Gustafson, 1991; Kramer and Feinstein, 1981).

Results

The number of paired observations recommended by Cicchetti and Fliess (cited in Fliess, 1981) was $n>3k^2 = 48$, where k is the number of categories, was obtained for all items except pruritus. Consequently, the results for pruritus should be interpreted with caution. Where examination of item contingency tables revealed a frequency of only one, this single event was recoded as missing data. Seventeen items had scores at three levels, "none". "mild" and "moderate". The computed kappas are provided in Table 1.

Kappa values, in absolute terms, for the dichotomous items varied from "fair" (.39) to "substantial" (.78) reproducibility. These kappas were statistically different from chance agreement. Of the pooled kappas, none reached an absolute value of .70 or greater. Even so, all (with the exception of trunk/scale) were estimated as significantly different from zero (alpha=.05). Agreement on lichenification was poorest with absolute values of pooled kappa ranging from .21 to .30. Agreement on erythema ranged from .32 to .45, for scale .41 to .50 (omitting trunk) and for excoriations .35 to .57. Agreement was higher for hands (range .45 to .52) and least for trunk (.13 to .38) omitting lichenification.

Discussion

Kappa results have indicated that there were statistically significant agreements between the doctors when rating features of AD on the ADAM measure. The absolute values of the kappas, however, indicated that agreement was far from optimal. Moreover, there were some serious problems with agreement with scale on the trunk and lichenification on any site. The skin on the trunk could act as a "baseline" in that it could be the site expected to be the most "normal". If this is the case the poorer agreement could reflect differences of opinion of what constitutes "normally" dry or scaly skin. Moreover, agreement on "severe" AD was not found.

Two thirds of the sample had been seen previously by clinic doctors. AD of familiar patients may be assessed with a bias. Bias and prevalence may reduce rather than increase the reliability of the kappas (Byrt et al., 1993). To the extent to which the scales were operationally defined, however, these response biases should have been diminished (Streiner and Norman, 1989).

Conclusion

The reliability study of the ADAM measure produced results consistent with agreement levels reported in the medical research literature (Sackett et al., 1991). Yet, agreements were far from optimal and imply that operational definitions of grades of severity of AD were not well-enough defined. Moreover, doctors had greater agreement on "mild" code levels than on "moderate" and "severe" code levels. It may be that a non-normal distribution of severity of AD represents a "true" reflection of presenting patients rather than being a statistical aberration or "noise". Another approach to the matter of variation (not across codes levels) but among raters is taken by Linacre, Englehard, Tatum and Myford (1994) and others. They argued (and demonstrated) that analyses should take into account the variations in severity of ratings among raters. However, to do this, sophisticated software and large sample sizes are required which are beyond the scope of most clinical studies. In an argument against their approach, medical research requires exact and agreed upon ratings as these are essential for diagnoses and as bases for interventions, for example, admission to hospital, surgery or laser therapy.

In this study, operational definitions for the ADAM measure were provided by clinicians. The SCORAD system too has clinical definitions with accompanying photographs. However, in the SCORAD system, there was no direct relationship between these definitions and mathematical basis for the severity measures. However, as a basis for improving results of agreement studies and for clinical training, a further study should be conducted to develop operational definitions of grades of severity that are directly related to the severity measure.

This further study will be a Partial Credit analysis. The first assumption for the application of Partial Credit modeling has been met, that is the ADAM measure has been shown to be uni-dimensional. While this conclusion may be subject to dispute (refer to McDonald, 1983; Hattie, 1985), Partial Credit model as developed by Adams and Khoo (1993) does have an alternative measure of uni-dimensionality within it and provides a number of indices to estimate goodness-of-fit of items and cases. Adams and Khoo (1993) provided for ordinal data and also for log transformations to accommodate non-normally distributed data. The next issue of the journal will report results of this Partial Credit analysis and how these results were used to derive operational definitions of clinical descriptions or "word pictures" of adversity.

Acknowledgements

The authors would like to acknowledge the assistance of Ted Byrt and Janet Bishop, Clinical Epidemiology and Biostatistics Unit, Royal Children's Hospital, Melbourne, Australia. This project was partly funded by the Royal Children's Hospital Research Foundation.

References

Agner, T. and Serup, J. (1990). Sodium lauryl sulphate for irritant patch testing-adoe-response study using bioengineering methods for determination of skin irritation. *J Investig Dermatol., 95*, 543-547.

Allen, K. and Harris, C. (1966). Elimination of a child's excessive scratching by training mothers in reinforcement procedures. *Behav Research Therapy, 4*, 79-84.

Bahmer, F.A., Schafer, J. and Schubert, H.J. (1991). Quantification of the extent and the severity of atopic dermatitis; the ADASI score. *Arch Dermatol., 127*, 1239-1240.

Bahmer, F.A. and Schafer, J. (1992). Die Behandlung der atopischen Dermatitis mit Borretsch-samen-Ol (Glandol)-eine Zeitreihenenalytische Studie. *I Praxis. 60*, 199-202.

Byrt, T., Bishop, J. and Carlin, J.B. (1993). Bias, prevalence and Kappa. *J Clin Epidemiol. 46*, 423-429.

Costa, C., Rilliet, A., Nicolet, M. and Saurat, J.H. (1989). Scoring Atopic Dermatitis: The simpler the better? *Acta Dermatol Venereol (Stockh.), 1* (69), 41-45.

Devlin, J. and David, T.J. (1992). Tartrazine in atopic eczema. *Arch. Dis. Childhood, 67*, 709-711.

European Task Force on Atopic Dermatitis. (1993). Severity scoring of Atopic Dermatitis: The SCORAD index. *Dermatology, 186*, 23-31.

Ewing, C.I., Gibbs, A.C.C., Ashcroft, C. and David, T.J. (1991). Failure of oral zinc supplementation in atopic eczema. *Eur J Clin Nutrition, 45*, 507-510.

Fliess, J.L. (1981). Statistical measures for rates and proportions. (2nd ed.). New York: Wiley.

Gil, K.M., Keefe, A., Sampson, H.A., McCaskill, C.C., Rodin, J. and Crisson, J.E. (1988). Direct observation of scratching behaviour in children with atopic dermatitis. *Behav Therapy, 19*, 213-227.

Gustafson, T.L. (1991). True Epistat Statistical Package. Richardson, Texas: Epistat Services.

Hall, D.E., Lynn, J.M., Altieri, J. and Segers, V.D. (1987). Inter-intrajudge reli-

ability of the stuttering severity instrument. *J Fluency Dis. 12*, 167-173.

Hanifin, J..M. and Lobitz, W.C. (1977). Newer concepts of Atopic Dermatitis. *Arch Dermatol., 113*, 663-670.

Hanifin, J.M. (1989). Standardized gradings of subjects for clinical research studies in Atopic Dermatitis. *Acta Dermatol Venereol (Suppl.). 144*, 28-30.

Jemec, G.B. and Wulf, H.C. (1997). The applicability of clinical scoring systems: SCORAD and PASI in psyoriasis and atopic dermatitis. *Acta Derm. Venereol., 77* (5), 392-3.

Kline, P.A. (1986). Handbook of test construction: introduction to psychometric design. Cambridge: Methuen.

Koblenzer, C.S. (1989). Psychocutaneous disease. Grune and Stratton Inc. Harcourt Brace Jovanovich:

Kramer, M.S. and Feinstein, A.F. (1981). The biostatistis of concordance. *Clin Pharmacol Therap. 29*, 111-123.

Kunz, B., Oranje, A.P., Labr'eze, L., Stalder, J.F., Ring, J. and Taib, A. (1997). Clinical validation and guidelines for the SCORAD index: Consensus report of the Eurpoean Task Force on Atopic Dermatitis. *Dermatology, 195*(1), 10-19.

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

Noren, P. and Melin, L. (1989). The effect of combined topical steroids and habit-reversal treatment in patients with atopic dermatitis. *Brit J Dermatol. 125*, 1063-8.

Rajka, G. and Langeland, H. (1989). Grading the severity of Atopic Dermatitis *Acta Dermatol Venereol (Stockh) Suppl. 144*, 13-14.

Rajka, G. (1986). Natural history and clinical manifestations of Atopic Dermatitis. *Clin Rev Allergy. 4*, 253-262.

Sackett, D.L., Haynes, R.B., Guyatt, G.H. and Tugwell, P.(1991). Clinical epidemiology: A basic science for clinical medicine. 2nd Edition. Boston: Little, Brown and Company.

Spiteri, M.A., Cook, D.G. and Clarke, S.W. (1988). Reliability of eliciting physical signs in examination of the chest. *The Lancet, 6*, 873-875.

Sprikkelman, A.B., Tupker, R.A., Burgerhof, H., Schouten, J.P., Brand, P.L., Heymans, H.S. and vanAalderen, W.M. (1997). Severity scoring of atopic dermatitis: A comparison of three scoring systems. *Allergy, 52*(9), 944-9.

Streiner, D.L. and Norman, G.R. (1989). Health measurement scales: A practical guide to their development and use. New York: Oxford University Press.

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

JOURNAL OF OUTCOME MEASUREMENT®, *3*(1), 35–52
Copyright© 1999, Rehabilitation Foundation, Inc.

Competency Gradient for Child-Parent Centers

Nikolaus Bezruczko
Chicago, Illinois

This report describes an implementation of the Rasch model during the longitudinal evaluation of a federally-funded early childhood preschool intervention program. An item bank is described for operationally defining a psychosocial construct called community life-skills competency, an expected teenage outcome of the preschool intervention. This analysis examined the position of teenage students on this scale structure, and investigated a pattern of cognitive operations necessary for students to pass community life-skills test items. Then this scale structure was correlated with nationally standardized reading and math achievement scores, teacher ratings, and school records to assess its validity as a measure of the community-related outcome goal for this intervention. The results show a functional relationship between years of early intervention and magnitude of effect on the life-skills competency variable.

Requests for reprints should be sent to Nikolaus Bezruczko, 1524 E. 59th Street, Chicago, Illinois 60637, e-mail: nbezruczko@email.msn.com

This research applies Rasch measurement methodology during the evaluation of longitudinal outcomes for a federally-funded early childhood intervention program. Researchers generally agree that the Chicago Child-Parent Centers[1] have long term school benefits for urban, disadvantaged children, and some research suggests more extended societal benefits (for a discussion of benefits see Conrad and Easch, 1983; Fuerst and Fuerst, 1993; Reynolds, Mavrogenes, Hagemann, and Bezruczko, 1993; Reynolds, 1989; Reynolds, 1994). Methods in these studies, however, usually only compare nationally standardized student achievement scores between groups, because objective measures of community-related student performance are extremely rare. These studies tend not to examine program outcomes on variables representing student success at completing socioeconomic transactions in the community, and they generally do not provide insight into the mental operations or skill structures underlying children's performance on school or community-based outcome variables. Consequently, the typical evaluation of a federally-funded social and especially educational program is not particularly revealing concerning positive social outcomes or specific program weaknesses.

This research first presents a theoretical context to introduce the concept of community life-skills competency, an intended program outcome of this early childhood intervention for socioeconomically disadvantaged African American children. Then an empirical variable is constructed to measure student performance in this context. After identifying items that reproduce the variation of student performance on a community life-skills variable, this analysis carefully examines these items to infer cognitive structures underlying successful performance on this variable. Because the program effects are longitudinal in their range extending from age three to grade 8, a functional relationship is finally examined between years of enrollment in the intervention program and children's life-skills development.

The life-skills competency model presented in this research provides an underlying construct for this Rasch analysis. Then through the development of a multiple-choice item bank, a variable is isolated which reveals children's development in this life-skills model. Finally, an application of BIGSTEPS software (Wright and Linacre, 1992) estimates empirical parameters for the children and the items. The unidimensional and objective empirical structure revealed by this analysis is then cross-validated with survey data, standardized student achievement scores, teacher ratings, as well as delinquency and school adjustment informa-

tion to establish its plausibility as an important aspect of a more general psychosocial construct, social competency.

This study addresses the following questions:

1. Can a complex, multidimensional psychosocial construct such as life-skills competency be usefully parameterized with the one-parameter Rasch model? If so, how do children who systematically differ in early intervention experiences differ in their placement on this structure?

2. Can this structure provide insight into the cognitive dynamics underlying a more general social competency construct?

Conceptual model for life-skills competency. The conceptual model underlying life-skills competency in this research is a hypothetical series of concentric spheres centered on the developing child. Each sphere defines a progressively more complex sociocultural context for the child beginning with family interactions in the home extending into the neighborhood and community (see Bronfenbrenner, 1977a, 1977b, 1979; Bronfenbrenner and Morris, 1998; Lewin, 1931, 1935). (Presumably, this structure extends into metropolitan area and nation, as well.) Each sphere has requirements concerning physical mobility, intellectual ability, and language proficiency and expects conformity to particular social rules and conventions. Each sphere also consists of various socioeconomic transactions which increase in their difficulty, hence their challenges to a person's social competency increase as one moves away from the center. A young child, for example, is expected to show reasonable competence around the home but is not expected to demonstrate considerable competence in the community. However, as children mature and gain experience with socioeconomic transactions hence acquire skills and knowledge essential for their success in the community, their competence is expected to increase. Previous studies show that children's performance on this structure is influenced by education, family characteristics and socioeconomic background, as well as personal attitudes and interests, inherited mental aptitudes, and cumulative life experiences.

Figure 1 presents a diagram of this hierarchical model with a hypothetical unidimensional competency variable bisecting it. In this example, the variable is demarcated by community-based "real life" skills and knowledge that presumably correspond to a particular sphere of social competency. In order to reproduce the underlying dynamics of children growing and developing through this structure, skills and knowledge are expected to become increasingly difficult.

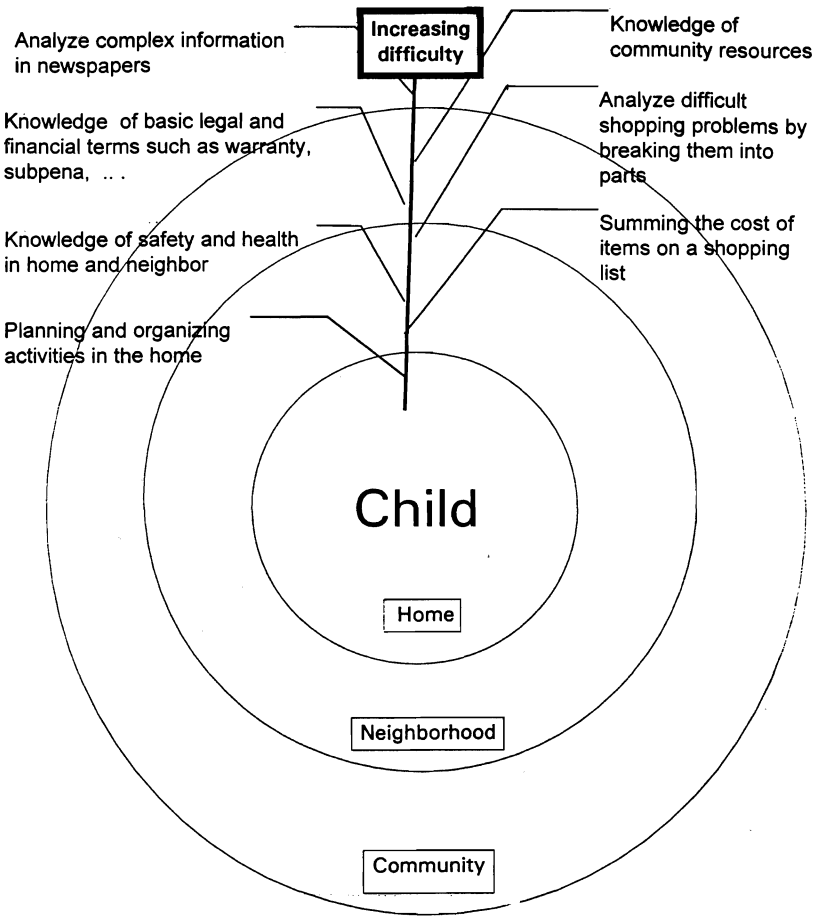


Figure 1. A contextual model for life-skills competency

Item bank design. A convenient and effective method for operationally defining the discrete challenges in a structure such as the competency model described above is to design a test matrix and write items for an item bank. An idealized property of items in a calibrated item bank is "specific objectivity" which represents the convergence of many similar items on a single quantitative construct defined by person ability and item difficulty (Rasch, 1960/1980). In practice, this idea is realized through the statistical invariance of the difficulty parameter. Items not having this property fluctuate in their difficulty depending on the sample being tested which limits their usefulness as measuring units. Consequently, a major responsibility of the program evaluator is to verify their invariance by examining difficulty estimates, comparing them with previous administrations and between groups whenever the items are administered.

Anchor items are a key element in an item bank, because all forms developed from the item bank are linked together by a set of anchor items (see Wright and Stone, 1979). While forms may change from year to year, or perhaps several forms are developed for a particular year, a small set of items in each form is identical. Figure 2 shows anchor items linking annual test forms to each other and back to a life-skills item bank.

Method

Sample. The sample is 828 eighth grade students in the Chicago Longitudinal Study of Child-Parent Centers [2]. These are socioeconomically disadvantaged African American children in the Chicago Public Schools who enrolled in a federally-funded Child-Parent Center for preschool and kindergarten in 1985 and 1986. Because intervention for some children continued through first, second, and third grade, length of intervention defines six discrete groups in the sample. The sample is equally divided into boys and girls.

Data. Community life-skills competency was measured by a 63 item multiple-choice test called the Minimum Proficiency Skills Test [3] (MPST; Bezruczko and Reynolds, 1987; Reynolds and Bezruczko, 1989) derived annually by the Chicago Public Schools, Department of Research and Evaluation from a calibrated item bank. This form (including 15 anchor items, five per subscale) is based on a test matrix consisting of three subscales (language arts, problem solving, and computation) that is applied in seven neighborhood- and community-related content areas (health, communication, finances, transportation, government, community re-

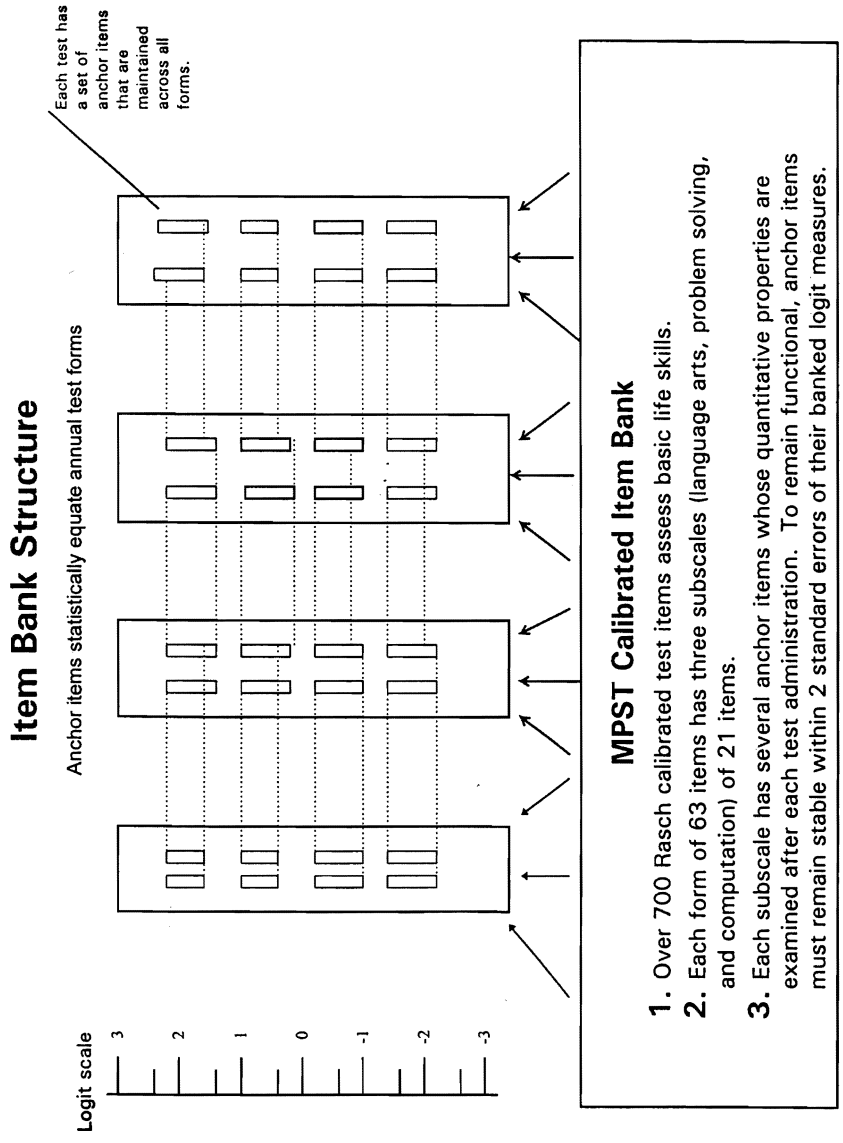


Figure 2. Item bank structure: In this research, an annual form of 63 items assessing basic life-skills competency was linked by 15 items to an item bank of approximately 700 items.

sources, and occupations). Unlike academic achievement test items, MPST items only present students with common socioeconomic transactions in their neighborhoods and communities [4]. In order to ensure the MPST is not a reading test, item readability does not exceed Grade 3. Annual alpha reliability is consistently above .85.

Other data collected include reading and math scores from the Iowa Tests of Basic Skills (ITBS: Hoover, Hieronymus, Frisby, and Dunbar, 1993), a survey of student attitude toward school and learning, and teacher ratings of student achievement. In addition, computerized school office administrative records were examined for student attendance, drop out, and delinquency information. A group of these students also provided writing compositions and were interviewed.

Procedure. The Chicago Public Schools, Bureau of Student Testing supervised the administration of the MPST and ITBS to all students during the second semester of their eighth grade. Surveys, ratings, administrative file searches, writing sample, and interview were collected by the Chicago Longitudinal Study.

Analysis. BIGSTEPS computer software estimated item and person parameters and computed item and person fit values (Wright and Linacre, 1992).

Results

Empirical variable. Figure 3 presents a plot of the students and items at their estimated positions on the calibrated competency variable. Students with lower scores and presumably less life-skills competence, as well as corresponding items, are located near the bottom of the structure. Higher scoring students and more difficult items appear at the top. The zero is arbitrarily set at the average item difficulty for the test. The average student ability for this sample is .92 logits (overall SD = .98) showing the majority scored above the middle of the test. A few of the students show ability higher than the test ceiling, and none of the students fall below the test floor. The obtained fit and reliability values generally support the measurement properties of this scale [5].

A dotted line on the left identifies 38 items or .49 logits as the "cut point" established by administrators and teachers for this test. All students at or above this point pass the test and are certified for graduation from the Chicago Public Schools. Students below this point fail the test and may take it again.

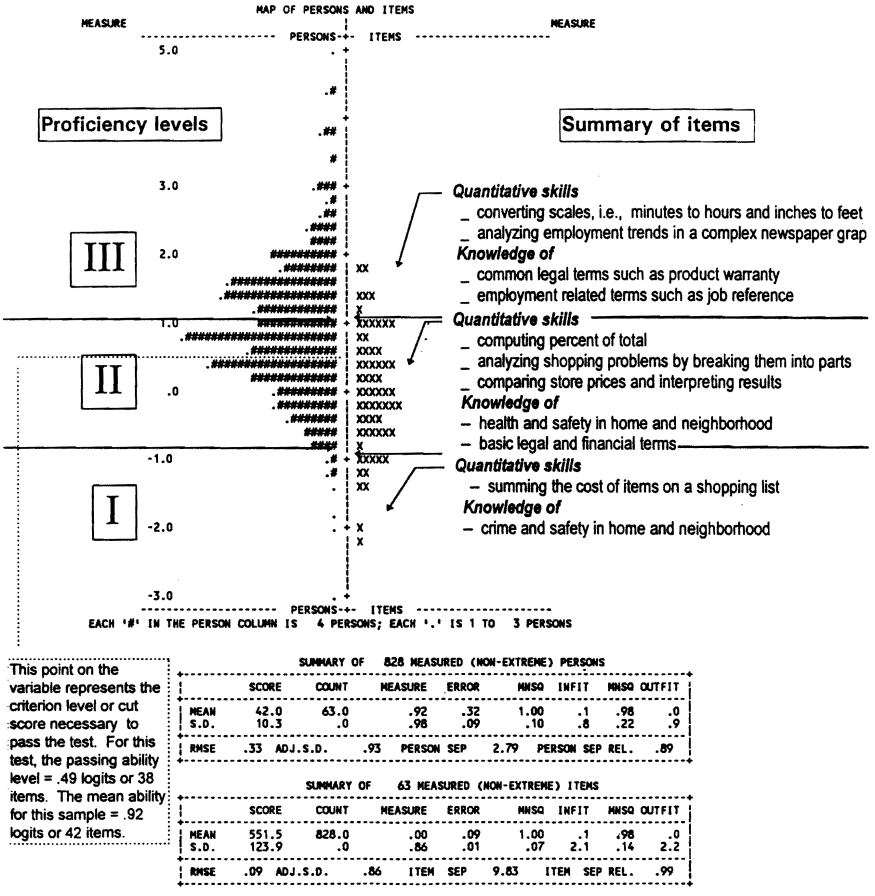


Figure 3. An assessment of life-skills competency. Level I items tend to require only simple arithmetic. Level II also requires simple arithmetic, but, in addition, requires skills like dividing a given problem into parts and reorganizing the parts to reach an answer. Level II items also put much stronger emphasis on analyzing and comparing before completing an arithmetic procedure. Level III differs from II by requiring some transformation of information before reaching an answer. Also, the arithmetic in several Level III questions is more complicated involving division and percentages.

The results show that among the easiest items is knowledge about filling in a job form. A more difficult item requires summing a group of numbers on a grocery shopping list. Among the most difficult items, one asks for knowledge about product warranty. According to these results, comprehension of legal terms such as "loan interest" and "subpoena" are at the middle of the test, and any student who passes the test has better than 50 percent probability of passing these particular items. The item distribution also appears loosely to form three clusters.

Cognitive operations. A qualitative analysis of the items suggests that two cognitive operations mediate performance for many of these items. They are the ability, first, to recall and understand common everyday knowledge, and, second, to conduct quantitative mental operations ranging from simple arithmetic to more complex analytical operations. These operations are required by many items and become more sophisticated as one progresses across the variable. The importance of these cognitive operations for the total score was tested by coding the items for these operations (0,1), as well as the number of words and the use of a graphic in the item, two other prominent differences among the items. Then the item calibrations were regressed on these codes. The results revealed the mental operations of simple recall and quantitative/analytical manipulation to account for over 30 percent of the variation in item difficulty ($R^2 = .31$; $F = 13.08$; $p<.001$).

Interpretation of test construct. The empirical analyses above ensures that the obtained item responses are consistent with the intentions of the measurement model and examining these results helps to identify irregularities in the measuring process. The analysis of cognitive operations helps to identify the mental skills required to pass the items which contributes to an evaluator's understanding of the cognitive dynamics underlying item responses. Likewise, a complementary analysis concentrates on the variation of person ability in this item structure. Figure 3, for example, shows these items separate the sample into three distinct ability levels suggesting these items define three levels of competence. (These levels were based on an estimated person separation = 2.79, and a mean standard error of measure = .09.) The lowest level of person ability or easy end of the variable (Level 1) is defined by simple knowledge items concerning crime and safety, as well as simple addition. A second more difficult level (Level II) requires knowledge of basic financial terms, and the mental ability to analyze and compare quantitative differences between physical objects. The highest level of competence on this scale

(Level III) requires knowledge of career-related terms and the quantitative ability to convert units of measurement. Consequently, these construct analyses suggest students systematically vary in their ability to analyze and reorganize information in a problem, transform and manipulate information in a problem, and in their capacity to grasp general knowledge. These competency levels are summarized below:

Level III

- converting scales of measurement
- analyzing a graph for trends and patterns
- knowledge of financial constructs such as warranty
- knowledge of career-related terms such as references

Level II

- computing percent of total
- analyzing and comparing concrete items
- knowledge of financial constructs such as loan and interest

Level I

- summing items on a list
- crime and safety in the neighborhood
- reading a job related form

These results provide insight into the behavioral structure of a competency variable whose validity as an aspect of social competence was further established with the following convergent and divergent analyses.

Construct validation. External validation of a test construct is usually completed by correlating measures with multiple data sources and examining their relationships. Items and criteria should converge and diverge in predicted directions given basic hypotheses about the conceptual model underlying the test variable. The MPST, for example, should correlate positively with standardized school achievement scores, as well as teacher ratings of student learning. The correlation with standardized achievement, however, should be moderate suggesting that school achievement and life-skills competence are unique constructs. Likewise, as a measure related to social competency, the MPST should correlate negatively with delinquency citations, school drop out status, and grade retention. (Students low on the variable may show serious social adjustment problems, and some students may eventually require adjudication.) Be-

cause socially competent students should do better in school and have more positive expressions of self concept, MPST scores should correlate positively with teacher ratings of school attitude and motivation. Finally, content analyses of student writing compositions and interviews should provide further evidence of positive adjustment.

The empirical results show the MPST to have:

- high positive correlations with student attitude and teacher ratings of learning and attitudes toward school (>.90).
- only moderate correlation with ITBS reading and math achievement scores (both approximately .70) accounting for only 50 percent of the MPST test variance.
- significant low negative correlations with delinquency records, grade retention, school attendance, and school drop out.
- significant low positive correlations with student attitudes toward future life expressed in written essays and interviews. Students, for example, who scored high on the MPST tend not only to have future goals but express specific ideas about how they see themselves in the future.

These results provide substantial evidence that the empirical structure provided by the Rasch analysis in fact has significant practical implications. (Students placing higher on the structure perform better in school and presumably will adapt better to adult life.) Further analyses of this structure established that length of participation in the CPCs is functionally related to children's position on the calibrated competency structure. These results are presented below.

Group differences. Finding a systematic relationship between duration of CPC intervention and student position on the competency variable would be important to program evaluators because of its significant policy implications. Consequently, six mutually exclusive groups were defined by the length of intervention. Students who received only one year of CPC preschool were classified CPC 1, while students who received two years of preschool, a kindergarten intervention, and three years of elementary school intervention were classified CPC 6. The results in Figure 4 show not only an ordered relationship between number of years of intervention, but any amount of intervention puts a child very near or above the minimum competency cut point on the MPST. This result has extraordinary implications for policy makers because approximately 40 per-

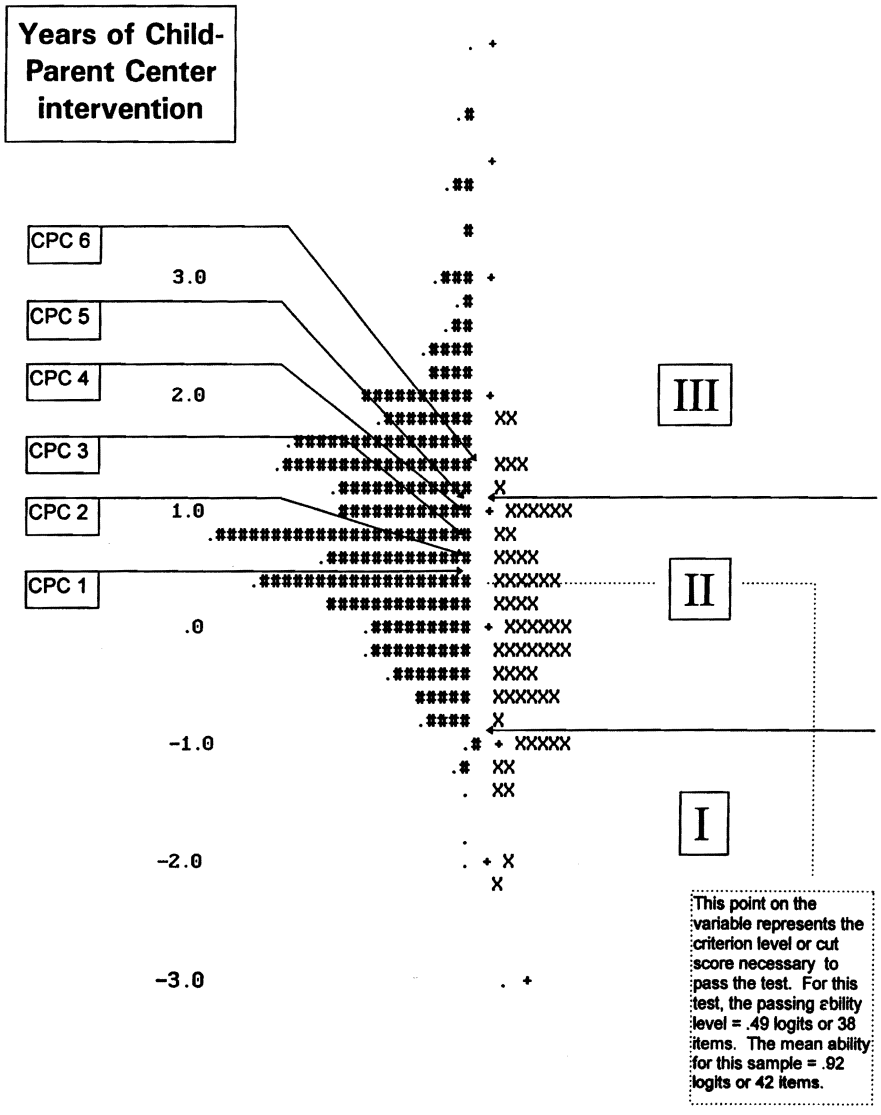


Figure 4. Relationship between length of intervention and position on the life-skills competency variable.

cent of the overall Chicago Public School eight grade population fall below the cut point hence fail the test on the first attempt, even though the overall CPS student population is much less disadvantaged than this sample. CPC 5 and CPC 6 are especially interesting because their performances cross into Level III of the variable suggesting that students who received intervention through second and third grade show significant qualitative differences in their cognitive functioning. Their grasp of common knowledge and their facility with life-skills quantitative operations are at the highest levels of the variable. Table 1 presents a comparison of this CPC cohort on this competency variable [5].

Table 1

Comparison of CPC groups on the MPST

CPC group	Raw score	logits	Effect size [a]	N
6	44.6	1.19	.71	71
5	43	1.00	.52	248
4	42	.97	.47	147
3	40.9	.80	.32	184
2	38.4	.63	.14	118
1	37.4	.58	.09	19

Note: The effect size was computed by dividing the difference between the passing criterion (.49 logits) and the average group ability by the overall person measure standard deviation (.98). These logit values were estimated by BIGSTEPS software (Wright and Linacre, 1992).

[a] The t-values for these effect sizes are 6.0, 8.2, 5.7, 4.3, 1.5, and 0.4, respectively. Their probability of occurring by chance is < .001.

Discussion

Large federally-funded social programs such as Head Start or the Educational Consolidation and Improvement Act (commonly referred to as Title 1) present special problems to government agencies because evaluations of these programs frequently do not yield results with practical public policy implications. Head Start, for example, has been examined for over 30 years by over 1000 empirical studies, yet researchers and policy makers still can not agree on its program outcomes (see Currie and Thomas, 1997). While I agree criticizing these studies retrospectively is un-

fair (many of them suffer from design flaws), all of them implemented measuring methods that limit the usefulness of their evaluations.

The methodology and results reported here offer program evaluators new tools to establish when social and educational programs succeed or fail, and these methods offer evaluators the unusual capacity to understand specific cognitive and social effects of these programs. Although inferring the underlying dynamics of scale structures from the position of calibrated items and measured persons on a scale is a well-established feature of Rasch analysis (Wright and Masters, 1982), it is rarely exploited in evaluations of federally-funded programs.

Concerning an important question motivating this research, we can now conclude that the complexity of a social phenomena such as life-skills competence has little bearing on the usefulness of the unidimensional Rasch model. The complex, multidimensional, and hierarchical concept presented here for life-skills competency was simply reconstructed in theoretical terms that require linear relationships among objective scale units. After this formulation, the interpretation of empirical relationships between persons and items on this structure produced new insights into how disadvantaged children become competent and worthwhile speculation concerning effective public policies.

The results from this Rasch analysis provide many useful insights into the outcomes of early childhood intervention that, hitherto, have not been revealed by conventional approaches to program evaluation. The emphasis on a theoretical socioeconomic context, arguably more important than classroom achievement outcomes, and directly measuring performance in this context revealed a CPC performance gradient on a measure of community life-skills competence. The practical importance of these analyses suggests that over time, the CPC may instill particular attitudes and develop specific mental capacities in young children that will benefit them as adults.

Although useful, these results show that the measurement of life-skills competency in this study depended primarily on students' capacity to recall common knowledge and verbally manipulate quantitative information in the social context of neighborhoods and communities. While important, a thorough evaluation of program outcomes probably requires a much more comprehensive response structure operationally defined by many more cognitive operations than identified for the MPST. The methods presented here were successful in measuring an important social out-

come for the CPC, but these results show the implementation of Rasch methodology was too restricted to convince policy makers of their importance.

Finally, these results demonstrate important methodological advantages gained from implementing an objective measurement model to evaluate program outcomes. Parameter invariance, as well as estimates of their precision, helped to establish an analytical framework that is stable and reproducible. Hence, the evaluation is not only descriptive, but provides a practical basis for predicting program benefits that will be realized after these students become adults. Likewise, linearization of the test scores provided important insight into a relationship between the length of intervention and program effectiveness. The subtlety of this relationship, despite its importance to policy makers, is not possible to identify with any precision from simple group comparisons of total test scores.

Acknowledgments

This research was supported by a grant from the National Institute for Mental Health. Portions of this study were presented at the National Meeting of the American Educational Research Association, Chicago, 1997. I sincerely appreciate the permission of the Chicago Public Schools, Department of Student Testing to analyze the Minimum Proficiency Skills Test scores. My deepest gratitude goes to Arthur J. Reynolds and the Chicago Longitudinal Study of Child-Parent Centers for permitting me to analyze the validation data reported here. This research would not have been possible without his cooperation. Finally, I am grateful to John Panontin for his comments of an earlier draft of this report.

References

Bezruczko, N. (1998). *Neighborhood and school context for the Chicago Longitudinal Study of Child-Parent Centers.* Unpublished report. Madison, Wisconsin: University of Wisconsin.

Bezruczko, N., and Reynolds, A. J. (1987). *Minimum proficiency skills test: 1987 item pilot report.* Chicago: Chicago Board of Education, Department of Research and Evaluation.

Bronfenbrenner, U. (1977a). Toward an experimental ecology of human development. *American Psychologist, 32*, 513-531.

Bronfenbrenner, U. (1977b). Lewinian space and ecological substance. *Journal of Social Issues, 33*, 199-213.

Bronfenbrenner, U. (1979). *The ecology of human development.* Cambridge, MA: Harvard University Press.

Bronfenbrenner, U. and Morris, P. (1998). Ecological processes of development. In W. Damon (Ed.) *Handbook of Child Psychology: Theoretical issues* (Vol. 1). New York: Wiley.

Conrad, K. J., and Easch, M. J. (1983). Measuring implementation and multiple outcomes in a Child-Parent center compensatory education program. *American Educational Research Journal, 20,* 221-236.

Currie, J., and Thomas, D. (1997). Can Head Start lead to long term gains in cognition after all? *SRCD Newsletter, 40,* 3-5.

Fuerst, J. S., and Fuerst, D. (1993). Chicago experience with an early childhood program. The special case of the Child-Parent Center program. *Urban Education, 28,* 69-96.

Hoover, H. D., Hieronymus, A. N., Frisby, P., and Dunbar, S. B. (1993). *Iowa Tests of Basic Skills: Primary battery.* Chicago: Riverside.

Lewin, K. (1931). Environmental forces in child behavior and development. In (C. Murchison, ed.) *A handbook of child psychology.* Worcester, MA: Clark University Press.

Lewin, K. (1935). *A dynamic theory of personality.* New York: McGraw-Hill.

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

Reynolds, A. J. (1989). A structural model of first-grade outcomes for an urban, low socioeconomic status, minority population. *Journal of Educational Psychology, 81,* 594-603.

Reynolds, A. J. (1994). Effects of a preschool plus follow-on intervention for children at risk. *Developmental Psychology, 30,* 787-804.

Reynolds, A. J. (1999). *Success in Early Intervention: The Chicago Child-Parent Centers and Youth Through Age 15.* Washington, DC: American Psychological Association.

Reynolds, A. J., and Bezruczko, N. (1989). Assessing the construct validity of a life-skills competency test. *Educational and Psychological Measurement, 49,* 183-193.

Reynolds, A. J., Mavrogenes, M., Hagemann, M., and Bezruczko, N. (1993). *Schools, families, and children: Sixth-grade results from the longitudinal study of children at risk.* Washington DC: U.S. Department of Education.

Reynolds, A. J., Mavrogenes, N., Bezruczko, N., and Hagemann, M. (1996). Cognitive and family-support mediators of preschool effectiveness: A confirmatory analysis. *Child Development, 67,* 1119-1140.

Wright, B. D., and Linacre, J. M. (1992). *BIGSTEPS.* Chicago: MESA Press.

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

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

Endnotes

[1] Child-Parent Centers were established in Chicago in the 1960s to better prepare socioeconomically disadvantaged, urban African American children for elementary school. (For a description of these Title I neighborhoods see Bezruczko, 1998.) The primary intervention strategies implemented in a typical Child-Parent Center preschool or kindergarten classroom are low teacher-child ratio, a teacher aide, a supervising teacher, extra supplies, and a mandatory parent training program. In some schools this intervention continues in the elementary school through first, second, and third grades. (The basic program is a year of preschool and kindergarten.) The program currently serves approximately 3,000 children in 24 centers at an annual cost of $20,000,000. An additional 1,500 children receive extended services in elementary schools.

[2] The Chicago Longitudinal Study (CLS) is an ongoing study of Child-Parent Center intervention effects for a cohort that graduated from Chicago Public School kindergartens in 1986 (Reynolds, 1989; Reynolds, Mavrogenes, Bezruczko, and Hagemann, 1996). Tracking these children annually through the Chicago Public Schools, CLS has documented higher reading and test scores, lower retention rates, less special education referrals, and more positive school attitudes for children from CPCs. This sample, now in high school, is providing further evidence concerning the influence of early intervention on dropout, delinquency, and occupational goals.

[3] The MPST was designed and developed by the Chicago Public Schools in 1976-1978. Passing a life-skills competency test became a graduation requirement in the Chicago Public Schools in 1977-78, and the test was administered annually until 1997 when minimum competency testing was terminated in the Chicago Public Schools. Consolidation of an item bank began in 1983 and continued through 1987. Administered during the eighth grade, the passing criterion was 64 percent. Annually, approximately forty percent of the overall Chicago Public School eighth grade population failed the test on the first administration. (Anecdotal evidence suggests these students ultimately dropped out of school.) School policy permitted a student to continue taking the test

through high school until passing it. Approximately fifteen percent of the students failed the test three or more times.

[4] Because this test exclusively emphasizes community socioeconomic transactions, some research has referred to it as a "consumer skills" test (see Reynolds, 1999). While I do not object to this interpretation of its construct, it is overly narrow. These items cut across a wide range of community-based socioeconomic transactions of which consumer skills is only a subset.

[5] Average person meansquare = 1.00, SD = .10, and both person infit and outfit values are close to expectations. Average item meansquare = 1.00, SD = .07, and both infit and outfit values are close to expectations. Item infit and outfit SD values, however, are high suggesting that some items should be examined more closely for unstable quantitative properties. (Item infit and outfit SD = 2.1, and 2.2, respectively.) For this sample, person separation reliability = .89, and item separation reliability = .99.

JOURNAL OF OUTCOME MEASUREMENT®, 3(1), 53–70
Copyright© 1999, Rehabilitation Foundation, Inc.

Alternate Forms Reliability of the Assessment of Motor and Process Skills

Karen N. Kirkley

Anne G. Fisher

Colorado State University

The purpose of this study is to examine the alternate forms reliability of the AMPS (Assessment of Motor and Process Skills) (Fisher, 1997a) where alternate forms means different pairs of AMPS tasks. The participants for this study were persons selected from the AMPS database who had performed four AMPS tasks. The participants varied in age, gender, diagnosis, and level of assistance needed to live in the community. The AMPS was administered by trained and calibrated occupational therapists according to standardized procedures. The data for the 91 participants were subjected to 12 many-faceted Rasch analyses to generate ADL motor and ADL process ability measures for each task and each set of paired tasks. Repeated measures ANOVAs revealed no time effect across the four AMPS tasks. Pearson product moment correlations between Tasks 1 and 2 combined and Tasks 3 and 4 combined were $r = .91$ and $r = .86$ for the ADL motor and ADL process scales, respectively. Calculation of the standardized difference (z) revealed that no more than 8% of the participants had ADL motor or ADL process ability measures that differed significantly between observations once we accounted for real differences in a persons performance; 80% of the paired ADL motor and ADL process ability measures remained stable within ± 0.5 logits when the participants performed two tasks. The AMPS ADL motor and ADL process scales can be used reliably in clinical practice and for research purposes.

Requests for reprints should be sent to Anne G. Fisher, Department of Occupational Therapy, Occupational Therapy Building, Colorado State University, Fort Collins, CO 80523 e-mail: Afisher@cahs.Colostate.edu.

Introduction

The purpose of this study is to examine the alternate forms reliability of the AMPS (Assessment of Motor and Process Skills) (Fisher, 1997a). The AMPS is used by occupational therapists to evaluate a person's ability to perform personal and instrumental activities of daily living (ADL). Personal activities of daily living (PADL) include self care activities such as bathing, dressing, or grooming. Instrumental activities of daily living (IADL) include meal preparation and home maintenance tasks such as vacuuming, doing laundry, or bed making. The AMPS manual currently lists 56 IADL tasks and 7 PADL tasks as options available for the person tested to perform. These 63 tasks vary in the relative challenge they offer the client who is assessed.

Before initiating an AMPS observation, the rater interviews the person to identify a subset of three to five ADL tasks that are familiar and relevant to the person's daily routine, and which offer the person an appropriate challenge. From that subset, the person chooses which two or three tasks they wish to perform during their AMPS observation. Choice is an important feature of the AMPS because performance is maximized when people have the opportunity to choose tasks that they perceive as meaningful and relevant (Doble, 1988; Dickerson and Fisher, 1997).

Raters fully orient and familiarize each person with the test environment prior to each ADL task the person chooses to perform. The person and rater agree on the essential components of the task to be performed (e.g., a meat sandwich, cut in half, served on a plate), including specific details pertaining to which options the client will choose (e.g., the person chooses to use mustard, not mayonnaise or butter, and ham, not salami or cheese, when preparing a sandwich). The person remains free to perform the task in his or her usual manner within the standardized guidelines of the task (Fisher, 1997a). For example, some people spread mustard directly on the bread while others spread the mustard on the meat or cheese after placing the meat or cheese on the bread. While some raters may see this as odd, the person making the sandwich would not receive a reduced score unless the person's method was illogical (e.g., spreading the mustard on top of the second piece of bread after making the sandwich).

After the person is tested, the rater scores the person on 16 ADL motor items (e.g., Lifts, Reaches, Grips) and 20 ADL process items (e.g., Initiates, Chooses, Gathers) for each ADL task performed. The rater en-

ters these raw item scores into his or her personal copy of the AMPS computer scoring software to generate an ADL motor ability measure and an ADL process ability measure. Further descriptions of the AMPS have been reported elsewhere (Fisher, 1997b; Goto, Fisher, and Mayberry, 1996; Park, Fisher, and Velozo, 1994). The development o the item bank for linked AMPS observations has been described by Fisher (1997b).

The ADL motor and ADL process ability measures are calculated using many-faceted Rasch analysis (Fisher, 1993, 1994, 1997a, 1997b; Linacre, 1993). Rasch analysis converts ordinal raw scores into linear ability measures by means of logistic transformation. Specifically, the AMPS person ADL ability measures are expressed as "logistically transformed probability measures (logits), which are linear measures that can be placed on an abstract continuum of greater or lesser ability" (Bernspång and Fisher, 1995, p. 4). The higher the person's ability measures, the more likely the person does not evidence motor or process ADL skill deficits that disrupt the effort, efficiency, safety, or independence of his or her ADL task performance. Because the many-faceted Rasch model for the AMPS considers tasks and raters as facets (with tasks being calibrated in terms of their relative challenges and raters calibrated in terms of their relative severities), the many-faceted Rasch analysis is used to adjust the person's final ADL motor and ADL process ability measures to account for the challenge of the tasks the person performed and the severity of the rater who scored the task performance.

Two major assertions of the many-faceted Rasch model for the AMPS are that (a) easy ADL tasks are more likely to be less challenging for all people than are harder ADL tasks, and (b) all ADL tasks are more likely to be easier for more able people than they are for less able people. That is, the AMPS task calibration hierarchies (motor and process) are asserted to remain stable across people of varying abilities. If this assertion is true, the tasks included in the AMPS should (and do) demonstrate acceptable goodness-of-fit to the many-faceted Rasch model for the AMPS (Fisher, 1993, 1994, 1997a, 1997b; Linacre, 1993).

Since the AMPS tasks have demonstrated goodness-of-fit to the many-faceted Rasch model for the AMPS (Fisher, 1993, 1994, 1997a, 1997b), the adjustments for task challenge that occur when a rater uses his or her copy of the AMPS computer-scoring software should result in stable AMPS motor and process ADL ability measures across paired AMPS task performances (different combinations of two performed AMPS tasks).

This assertion can be conceptualized as a form of reliability, specifically reliability of alternate forms (Crocker and Algina, 1986).

"Reliability refers to the consistency of examinees relative performances over repeated administrations of the same test or parallel forms of the test" (Crocker and Algina, 1986, p. 127).

> The traditional formulation of test "reliability" can be derived from a "true score" model which assumes that the observed test score of each person can be resolved into two components: an unknowable true score and random error. Test reliability is defined as the portion of a sample's observed score variance SD^2 which is due to the samples's true score variance ST^2
>
> $$R = ST^2/SD^2 = 1 - (SE^2/SD^2)$$
>
> where the observed variance is partitioned into two components $SD^2 = ST^2 + SE^2$, and SE^2 is the error variance of the test, averaged over that sample (Wright and Masters, 1982, p. 113).

An advantage of Rasch analyses is that a direct estimate of the modelled standard errors (*SE*) are generated for each person. The *SE* tells us how precisely the ADL motor and ADL process ability measures of any person are estimated. The *SE* for each individual can be used to define a 95% confidence interval at ± 2 *SE*s (Wright and Stone, 1979). More specifically, the person ability measures for two alternate forms of a test may be plotted against each other in a scatterplot and the *SE* for each person can then be used to define control lines ± 2 *SE* from the diagonal identity line (Wright and Masters, 1982). The identity line represents a line along which equivalent test scores would be plotted. When the person ability measures remain stable over time, the plotted values fall along the diagonal identity line. Although some variation between tests is expected, 95% of the person ability measures should fall within a confidence interval defined by ± 2 *SE* (Wright and Stone, 1979). The advantage of the scatterplot method for evaluating alternate forms reliability is that individual scores falling outside of the control lines can be identified as being significantly different between each set of ADL task performances ($p \leq .05$). The person's data can then be examined to determine the source of the significant difference.

There is a need to compare the alternate forms reliability of the AMPS when a person's ADL motor and ADL process ability measures are based on the performance of only one AMPS task instead of the two AMPS

tasks that are typically performed. This has important implications for both research and practice. An AMPS observation generally consists of two or more ADL task performances (two being the most common). The scores from all AMPS task performances are combined to generate a single person ADL motor ability and a single ADL process ability measure for each AMPS observation .

Occasionally, a person will perform only one AMPS task because of fatigue, risk of injury, or time limitations. As with any test, reliability is increased when the number of items is increased (Crocker and Algina, 1986). By performing two tasks, the ADL motor ability measure is based on the raw scores for 32 items (16 x 2) rather than just 16, and the ADL process ability measure is based on the raw scores for 40 items (20 x 2) rather than just 20 items. It has been asserted, therefore, that by having a person perform two tasks rather than one, error is reduced and the estimated ADL motor and process ability measures predict more accurately the person's true ADL ability. This assertion is based on empirical reasoning, but has not been tested for statistical significance. The focus of this, study, therefore, was the examination of the stability of the AMPS ADL motor and process ability measures when using alternate forms, where alternate forms means the use of different AMPS tasks or different pairs of AMPS tasks. Specifically, we sought to answer the following questions: How reliable is the AMPS when two different ADL tasks are performed during each observation? How reliable is the AMPS when only one AMPS task is considered instead of two?

Methodology

Participants

The participants for this study were selected from the AMPS database which consisted, at the time of this study, of approximately 10,200 persons from twelve countries (United States, Canada, Israel, New Zealand, Australia, Norway, Sweden, Denmark, Holland, United Kingdom Japan, Hong Kong) who had been scored by trained and calibrated AMPS raters. Three categories of participants were then eliminated from consideration for inclusion in the analysis. Participants who had been co-scored by more than 10 raters for purposes of calibration were eliminated due to potential bias related to multiple co-ratings. Well adult participants were eliminated because the AMPS was not designed to test healthy adults who do not demonstrate functional difficulty. Finally, approximately 5%

of the participants in the AMPS database had previously been eliminated because they had been identified as participants whose ability measures were associated with rater scoring error. Raters scoring error occurs when raters allow clients to perform talks of insufficient challenge. From the remaining participants, those who had performed four different tasks within a 7-day period were selected as candidates for inclusion in this study. A total of 91 participants who varied in age, gender, diagnosis, and level of assistance needed to live in the community were included in the study (Table 1).

Table 1

Participant Demographics for Age, Diagnosis, Functional Level, Gender, and Ethnicity

Age (years)	
M	64.6
SD	18.2
Range	17 to 90
Diagnosis	
Nondisabled older persons 60 years and above	16
Developmental disability	1
Dementia and memory impairment	5
Orthopedic/musculoskeletal	7
Psychiatric	8
Stroke	19
Other neurologic	12
Medical	5
Multiple diagnosis	18
Functional level	
Lives independently in the community	30
Needs minimum assistance to live independently in community	31
Needs moderate to maximum assistance to live independently in the community	30
Gender	
Male	31
Female	60
Ethnicity	
White	87
Black	3
Asian	1

Instrumentation and Procedures

The AMPS was administered by trained and calibrated occupational therapists according to standardized procedures (Fisher, 1997a). To become a trained AMPS rater, occupational therapists attended a 5-day AMPS training workshop. During the workshop, trainees co-scored videotaped observations of participants already in the AMPS database. After the course, they scored 10 additional participants to complete the calibration process. Their scores were analyzed using FACETS, a many-faceted Rasch measurement computer program (Linacre, 1993), to complete rater calibration (determine rater severity and reliability). Raters completing this process demonstrate high intra- and inter-rater reliability as evidenced by 95% of the raters demonstrating goodness-of-fit to the many-faceted Rasch model for the AMPS (Fisher, 1997a).

Data Analysis

The data for the 91 participants selected for this study were subjected to many-faceted Rasch analyses. Twelve many-faceted Rasch analyses were performed to generate the ADL motor and ADL process ability measures that subsequently were subjected to further statistical analyses. Task order was not randomized so that we could more closely examine

Table 2

Summary of the Many-Faceted Rasch Analyses Performed

Analysis	Task performances included
Motor Scale	
1	Task 1 only
2	Task 2 only
3	Task 3 only
4	Task 4 only
5	Task 1 & 2 combined
6	Task 3 & 4 combined
Process Scale	
1	Task 1 only
2	Task 2 only
3	Task 3 only
4	Task 4 only
5	Task 1 & 2 combined
6	Task 3 & 4 combined

the effects of time on the reliability of the AMPS ability measures. A summary of the Rasch analyses performed are shown in Table 2.

Results

Our primary research question addressed the alternate forms reliability of the AMPS—does the performance of two tasks during an AMPS observation result in the same ADL motor and ADL process ability measures as does the performance of two *different* tasks during another AMPS observation? Our secondary question pertained to the impact on reliability if people performed only one ADL task versus the usually performed two ADL tasks—how reliable is the AMPS when only one ADL task is performed?

To determine if there was a time effect across the four AMPS tasks, we performed two one-way repeated measures ANOVAs, one for the ADL motor scale and one for the ADL process scale. There was no significant overall time effect, $F(3,201) < 1.63, p > .05$. We then proceeded to examine alternate forms reliability using Pearson product moment correlation coefficients and standardized differences (z) (Table 3).

Table 3

Summary of Statistical Comparisons

No.	Scale	Task	Methods of comparison*
1	Motor	Task 1 vs. task 2 vs. task 3 vs. task 4	ANOVA
2	Process	Task 1 vs. task 2 vs. task 3 vs. task 4	ANOVA
3	Motor	Task 1 vs. task 2	Ppmc z
4	Process	Task 1 vs. task 2	Ppmc z
5	Motor	Task 1 & 2 vs. task 3 & 4	Ppmc z
6	Process	Task 1 & 2 vs. task 3 & 4	Ppmc z

* Ppmc = Pearson product moment correlation coefficients
z = standardized difference
ANOVA = Analysis of variance

Pearson product moment correlations between Task 1 and Task 2, and between Tasks 1 and 2 combined and Tasks 3 and 4 combined, are shown in Table 4. According to Hinkle, Wiersman, and Jurs (1988), correlations between .30 and .50 are low, correlations between .50 and .70 are moderate, correlations between .70 and .90 are high, and correlations

between .90 and 1.00 are very high. High positive to very high positive coefficients were found between Tasks 1 and 2 combined and Tasks 3 and 4 combined. When the ability measures for only Task 1 were compared to those for Task 2, the correlations were lower, but still within the high positive range.

Table 4

Pearson Product Moment Correlation Coefficients

Task	Motor	Process
Task 1 vs. task 2	.81	.71
Task 1 & 2 combined vs. task 3 & 4 combined	.91	.85

Scatterplots (Figures 1 through 4) are a visual representation of the relationship between two ability measures. Figures 1 and 2 display the person ADL ability measures for the AMPS motor scale. Figures 3 and 4 display the ADL ability measures for the AMPS process scale. Figures 1 and 3 show ADL ability measures derived from one task (Task 1 vs. Task 2) and Figures 2 and 4 show ADL ability measures derived from two tasks (Tasks 1 and 2 combined vs. Tasks 3 and 4 combined). The distribution of the ADL ability measures supports high positive relationships between paired ADL task performances.

Figures 1 through 4 also show control lines delimiting ± 2 *SE* from the diagonal identity line. On the motor scales, 4 participants (4.4%) had ADL ability measures that fell outside of the control lines when completing one task and 7 participants (7.7%) had paired ADL ability measures that fell outside of the control lines when completing two tasks. Twelve participants (13.2%) had paired ADL process ability measures that fell outside of the control lines when completing one task and 13 participants (14.3%) had paired ADL process ability measures that fell outside the control lines when completing two tasks. Calculation of the standardized difference (z) revealed that the same participants whose paired ability measures fell outside the control lines had ADL motor or ADL process ability measures that differed significantly between observations, $p \leq .05$. Since the standardized difference is based on the *SE*, a value that varies across individuals, it is useful in research, but is not of practical use in the clinical setting. Therefore, we calculated the percentage of paired ADL ability measures that did not differ beyond a variety of pre-specified ranges. As shown in Table 5, 80% of the paired ADL motor and ADL process

ability measures remained stable within ± 0.5 logits when the participants performed two tasks. When one task was performed, the proportion of scores that remained stable within ±0.5 logits dropped to approximately 67%.

Table 5

Percentage of Participants Whose Ability Measures Differed by Specified Logit Value

Scale and number of tasks	Logits			
	± .30	± .50	± .70	± .90
Motor 1 task	44%	69%	80%	88%
Motor 2 tasks	59%	80%	88%	93%
Process 1 task	48%	64%	76%	88%
Process 2 tasks	50%	81%	92%	97%

Discussion

The purpose of this study was to determine the alternate forms reliability of the AMPS when a person performs either one task or when the recommended two tasks are performed. This was accomplished by subjecting the ADL motor and process ability measures generated by Rasch analyses to traditional and nontraditional statistical analyses. Two of the traditional ways of verifying reliability are to (a) ascertain that two sets of measures are highly correlated (Pearson product moment correlation coefficient) or (b) verify that the mean ability measures do not differ significantly across time (*t* test or ANOVA). There are several limitations of traditional statistical approaches (e.g., correlations, *t* tests, ANOVAs). First, the error term is based on the overall sample *SE* and not the *SE* for each participant (overall group error vs. individual error). Furthermore, traditional statistics provide only an index of reliability for the overall group and are not able to target individual participants whose performances were unreliable. Finally, traditional statistics are known to be sample dependent and therefore are affected by the size and heterogeneity of the sample (Crocker and Algina, 1986; Wright and Masters, 1982; Wright and Stone, 1979). To overcome the limitations of traditional statistics, we have supplemented them with the use of the standardized difference (*z*), a statistical method that can identify individuals whose ADL ability measures differ significantly. The data can then be examined to identify the source of variance between two AMPS observations.

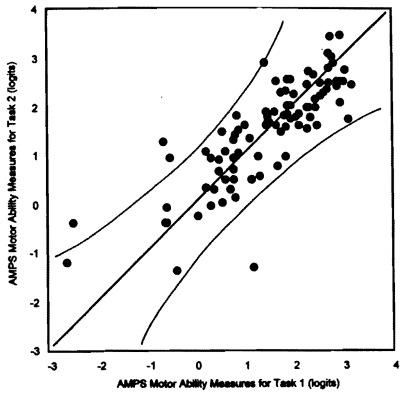


Figure 1. Scatterplot of the relationship of the person ADL ability measures for Task 1 vs. Task 2 - Motor

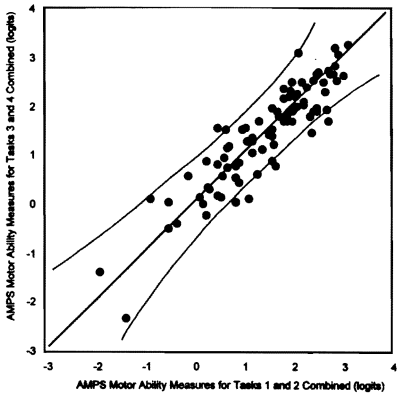


Figure 2. Scatterplot of the relationship of the person ADL ability measures for Tasks 1 & 2 combined vs. Tasks 3 & 4 combined - Motor

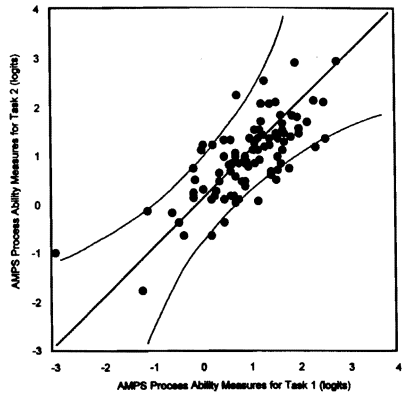


Figure 3. Scatterplot of the relationship of the person ADL ability measures for Task 1 vs. Task 2 - Process

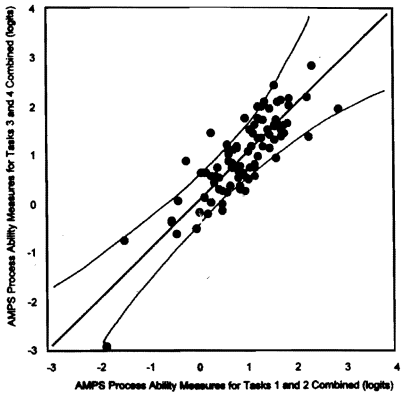


Figure 4. Scatterplot of the relationship of the person ADL ability measures for Tasks 1 & 2 combined vs. Tasks 3 & 4 combined - Process

Specifically, we computed the standardized difference for each individual to determine if his or her ADL ability measures differed significantly between paired AMPS observations. We then identified those participants who had a standardized difference above the critical value of 2.00 ($p \leq .05$). We expected 95% of the participants to have z values below 2.00.

The use of standardized difference is equivalent to examining the reliability of the AMPS by plotting the paired ADL person ability measures against each other and establishing a 95% confidence interval delimited by ± 2 *SE* (Figures 1 through 4). That is, the use of the scatterplot method creates the visual equivalent of the standardized difference. Once again, we would expect 95% of the plotted ability measures to fall within the control lines, $p \leq .05$.

When we compared the results of traditional and Rasch based methods, we found that traditional statistical methods revealed high alternate forms reliability of the AMPS, regardless of the number of tasks that were performed, but that the reliability coefficients attenuated slightly when the ADL motor or ADL process abilities for only one task were compared (Table 4). The ANOVAs also indicated high reliability as there were no significant differences among mean ADL motor or ADL process abilities for any of the four ADL tasks performed.

These results became more explicit when we compared them with the results of the Rasch-based standardized difference and the equivalent visual comparison provided by the scatterplot method. Visual analysis of the scatterplots, especially Figures 2 and 4, showed that the ability measures were closely grouped along the identity line, indicating stable ability measures when two tasks were performed. The ability measures fell further away from the identity line and in a more random pattern when only one task was completed (Figures 1 and 3). This indicated a greater amount of variance between ability measures when only one task was performed. The increased variance can also be seen visually by the increased distance between the 95% confidence interval control lines when only one task was performed. Additionally, the average *SE* increased from .30 to .45 on the ADL motor scale and from .23 to .35 on the ADL process scale when the ability measures were based on one task rather than two.

The fact that fewer of the paired ability measures fell outside of the control lines when Task 1 was compared to Task 2, than when Task 1 and 2 was compared to Task 3 and 4, is a result of increased estimation error (*SE*) when only one task was performed. Because the *SE* is larger when completing only one task, the ability measures can vary by a greater amount and still fall within the control band delimited by ± 2 *SE*. The larger error when only one task is performed is also reflected in the attenuated reliability estimates for one task shown in Table 4.

The increased *SE* and the decrease in reliability coefficients when completing only one task underscores the importance of having a person perform at least two AMPS tasks during a single AMPS observation. They also highlight the potentially misleading conclusions one can make based solely on traditional methods for examining reliability. That is, the relatively small attenuation of the reliability coefficients may not be recognized as being associated with such a large increase in the *SE*.

When we looked at the control bands delimited by ± 2 *SE*, and indicative of a 95% confidence interval, we expected no more than 5% of the paired ability measures to fall outside of the control lines. There were 4.4% to 7.7% of the participants whose plotted ability measures fell outside of the control lines on the ADL motor scale, and 13.2% to 14.3% whose plotted ability measures fell outside of the control lines on the ADL process scale. This did not meet the 5% criteria we had established indicating failure to meet statistically significant, $p \leq .05$, alternate forms reliability. However, clinical meaningfulness rather than statistical significance may be more useful to the occupational therapist who uses the AMPS. Clinical meaningfulness can be derived from Table 5. Occupational therapists who use the AMPS to evaluate intervention efficacy or who are concerned with the degree of confidence one can have in a client's ADL ability measures can feel *reasonably* confident that differences in ability measures greater than 0.5 logits are clinically meaningful if a person performed two tasks. However, if a person only performed one task, a similar level of confidence must be based on a difference of 0.7 logits or more. The use of 1 task performance for generating ability measures, therefore, results in a tool that is a less sensitive measure of change.

Another advantage of our using the standardized difference based on the individual *SE* is that it allowed us to identify those participants whose paired ADL abilities fell outside of the control lines. We could then attempt to determine the reasons for the variations in their estimated ADL abilities.

While we found no consistent pattern associated with rater, diagnosis, age, or task performed, we became aware of four potential reasons AMPS ability measures might vary. First, clinicians using the AMPS commonly report observing day to day, even hour to hour variations in ADL task performances among their clients, and the AMPS is sensitive enough to measure these changes. We were able to verify, with the original rater, that this was the case in 1 participant who had significant differ-

ences in her motor ability measures and 3 participants that had significant differences in their process ability measures. For example, one person had a decrease in their estimated ADL ability measures due to arthritis symptoms that were exacerbated by changing weather conditions. Another client with a brain injury was more agitated and disorganized on the first day he was tested than he was on the second day. These changes were associated with a clinically observable differences in the clients' ADL ability measures, and the AMPS was sensitive enough to detect these changes in a persons' performance. More importantly, this source of variance is not due to error, but the ability of the AMPS to detect real changes in performance. Careful decisions can be made regarding the advantages or disadvantages of assessing someone whose symptoms are exacerbated, or the value of documenting how a clients' ADL task performances vary over time. Moreover, when we account for participants who demonstrated real changes in their performance, only 7% of the participants on the ADL motor scale and 8% of the participants on the ADL process scale differed significantly. This is close to our expected 5% criteria.

A second reason AMPS ability measures vary is related to rater administration errors which include allowing the clients to perform poorly targeted AMPS tasks as well as rater scoring error. Clinical experience with the AMPS has revealed that some raters do not administer the AMPS correctly, allowing their clients to perform tasks that clearly do not offer an appropriate challenge. It is almost always the performance of a task that is too easy that is a problem, but occasionally tasks are offered that are too difficult. Both conditions result in unreliable measures.

There are several reasons why a rater may allow a person to perform a task that was too easy. The one that is of most concern is when the rater appears not to understand the importance of ensuring that the client performs appropriately challenging tasks. In other instances the problem appears to be more related to limitations of the AMPS rather than due to rater error. For example, in some instances men have refused to perform certain cooking tasks, stating "that is women's work." Currently there are no other more challenging task options. In other instances, people initially indicate that they are willing to perform certain tasks, but refuse to do so for an AMPS observation. In this case, clinicians have reported that they suspect that the clients are fearful that the extent of their disability (actual or perceived) may be "realized" if they perform one of the AMPS tasks that are more challenging. These are often persons still able to live

at home, but who fear their problem (e.g., memory loss) may result in "institutionalization." Finally, there is greater risk of a person performing a task of insufficient challenge when clients are expected to perform four different tasks, a situation that should occur only for research purposes (e.g., examination of the alternate forms reliability of the AMPS). To the extent that we could contact the original raters, we were able to determine that at least 2 participants on the AMPS process scale performed a task of insufficient challenge.

In clinical practice, the rater can be alerted that a task of insufficient challenge has been performed by monitoring the scores he or she assigns. When a rater awards raw item scores of 4 on most of the AMPS skill items on either scale, the rater should consider having the person perform an additional AMPS task of greater challenge to avoid error associated with performance of tasks that are not of sufficient challenge.

The rater can also be alerted that a problem may exist if the client's raw scores just varied noticeably between two tasks, but neither task was obviously too easy. In this case it is also advisable to have the person perform a third task of comparable challenge to the first two. Performance of a third task may help to clarify the reason for the difference in scores. For example, 3 participants (2 with rheumatoid arthritis and 1 with severe asthma) who varied on the ADL motor scale were person's whose performances were lower on tasks that required a greater amount of gross motor skills. That is, these three person's did better when they performed meal preparation tasks than when they performed house cleaning tasks (vacuuming, sweeping, changing sheets). Although this only occurred in 3% of the participants on the motor scale, it is a source of error that warrants further investigation.

Summary

Traditional statistics indicate that the AMPS is a reliable assessment tool. The reliability coefficients were high to extremely high (r = .91 for motor and r = .85 for process) when two tasks were completed. The results of the ANOVAs indicated that there was no significant difference in the mean ability measures over time. The use of the standardized difference identified no more than 8% of the participants who differed significantly due to error after we accounted for real differences in a persons performance on either the ADL motor scale or the ADL process scale. We also determined that we can have reasonable confidence that ability measures

will remain stable within ± 0.5 logits and that differences in a person's ability measures beyond that are likely a result of actual changes in the person's performance and not due to testing error. We have also determined that there is considerably more variation in ability measures when only one task is performed. The *SE* is increased and a similar level of confidence in the stability of measures must be based on ±0.7 or more logits when only one task is performed.

Conclusions

The AMPS ADL motor and ADL process scales can be used reliably in clinical practice and for research purposes. The use of reliable assessment tools is important since therapists use such instruments to make judgements about safety and independence, document progress of their clients, and make decisions regrading intervention. Although the AMPS has been shown to be reliable, this does not diminish the need for professional judgement on the part of the occupational therapist using the assessment. There are several factors, identified or confirmed through this research, that may decrease the reliability of a person's AMPS ADL ability measures. The therapist needs to be informed of the potential sources of error associated with the AMPS, and take the necessary precautions to assure that the results of the AMPS are truly representative of the person being assessed.

Recommendations

Although our results indicate that the AMPS is reliable, several possibilities exist for future research. Despite the fact that the AMPS tasks have been shown to fit the multi-faceted Rasch model for the AMPS (Fisher, 1993, 1994, 1997a, 1997b) we identified 3 participants that did better on tasks that are primarily fine motor tasks, and significantly worse on tasks that required more gross motor skills. Future research is needed to determine if some tasks emphasize certain skills over others causing their difficulty to change for some people with certain diagnoses.

We also identified poorly targeted AMPS tasks (tasks that were too easy for the client) as the most common issue associated with error in AMPS. Research is necessary to add more tasks to the AMPS. By adding more tasks, at all levels of difficulty, it would make it easier to target tasks that are of appropriate challenge to the client. While PADL tasks have been added to expand the lower end of the AMPS scales (Fisher, 1997b)

PADL data was not sufficient at the time of this research to determine the alternate forms reliability when PADL tasks were performed. Future research is necessary to determine the alternate forms reliability of the AMPS as new tasks are added.

In future research, we would recommend that raters include information about the person's affect or any condition that may cause fluctuations in the persons performance (e.g., exacerbations, mood changes). The raters of several participants that differed significantly were able to provide us with information that helped us identify the source of variance between ability measures. It would have been helpful to have this information on all clients in order to differentiate actual performance changes from measurement error.

Acknowledgement

This study was completed in partial fulfillment of the requirements of the Master of Science degree of the first author, Spring 1998, Colorado State University, Fort Collins, Colorado. It was supported in part by grant RO1-AG12345 from the National Institute on Aging, National Institutes of Health, US Department of Health and Human Services.

References

Bernspång, B., and Fisher, A.G. (1995) Validation of the Assessment of Motor and Process Skills for Use In Sweden. *Scandinavian Journal of Occupational Therapy, 2*, 3-9.

Crocker, L., and Algina, J. (1986) *Introduction to Classical and Modern Test Theory*. San Francisco: Holt, Rinehart and Winston.

Dickerson, A.E., and Fisher, A.G. (1997). The effects of familiarity of task and choice on the functional performance of young and old adults. *Psychology of Aging*, 12, 247-254.

Doble, S. (1988). Intrinsic motivation and clinical practice: The key to understanding the unmotivated client. *Canadian Journal of Occupational Therapy, 55, 75-81.*

Fisher, A.G. (1997a). *Assessment of Motor and Process Skills* (2nd ed.). Fort Collins, CO: Three Star Press.

Fisher, A.G. (1997b). Multifaceted measurement of daily life task performance: Conceptualizing a test of instrumental ADL and validating the addition of personal ADL tasks. *Physical Medicine and Rehabilitation: State of the Art Reviews, 11,* 289-303.

Fisher, A.G. (1994). Development of a functional assessment that adjusts ability measures for task simplicity and rater leniency. In M. Wislon (Ed.), *Objective measurement: Theory into practice*, (Vol 2, pp. 145-175). Norwood, NJ, Ablex.

Fisher, A.G. (1993). The assessment of IADL motor skills: An application of many-faceted Rasch analysis. *American Journal of Occupational Therapy, 47*, 319-338.

Goto, S., Fisher, A. G., and Mayberry, W. L. (1996). AMPS applied cross-culturally to the Japanese. *American Journal of Occupational Therapy, 50,* 798-806.

Hinkle, D.E., Wiersman, W., and Jurs, S.G. (1988). *Applied Statistics for the Behavioral Sciences (2nd ed)* Houghton Mifflin, Boston.

Linacre, J.M. (1993). *Many-faceted Rasch measurement.* Chicago: MESA.

Park, S., Fisher, A. G., and Velozo, C. A. (1994). Using the Assessment of Motor and Process Skills to compare occupational performance between clinic and home settings. *American Journal of Occupational Therapy, 48,* 697-709.

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.

JOURNAL OF OUTCOME MEASUREMENT®, *3*(1), 71–88
Copyright© 1999, Rehabilitation Foundation, Inc.

Teacher Receptivity to a System-Wide Change in a Centralized Education System: A Rasch Measurement Model Analysis

Russell F. Waugh
Edith Cowan University
Western Australia

The Education Department of Western Australia has implemented a new system called Student Outcome Statements, by trial in 1995/1996, then on a voluntary basis from 1997, with the intention of making it mandatory after 2001. The system describes, in order, the outcomes that students are expected to achieve in eight broad learning areas. The study has three aims. One, to create a scale for teacher receptivity to the use of Student Outcome Statements, based on eight orientations to receptivity: evaluative attitudes, behavior intentions, feelings towards Student Outcome Statements compared to the previous system, the benefits of the new system, support from significant others, alleviation of concerns, collaboration with other teachers, and involvement in decision-making. Two, to analyze the psychometric properties of the scale using the Extended Logistic Model of Rasch (Andrich, 1988; Rasch, 1960/1980) with the computer program RUMM (Andrich, Sheridan & Luo, 1997). Three, to provide advice to decision-makers about how better to implement the system of Student Outcome Statements.

Address correspondence to: Dr. Russell Waugh, Edith Cowan University, Pearson Street, Churchlands, Western Australia, 6018, e-mail: R.Waugh@cowan.edu.au

The Education Department of Western Australia (1996, 1994) trialed a new system in 1995/1996 called Student Outcome Statements. Teachers in government schools were expected to plan their teaching programs, prepare assessments of student achievement and report the progress of student achievement according to this system. The system described student outcomes in order from the beginning of primary school (6 years old) to the end of compulsory secondary school (15 years old). The outcomes reflect the knowledge, understandings, processes and skills considered essential for all students, across eight broad learning areas. These are: the Arts, English, Health and Physical Education, Languages other than English, Mathematics, Science, Society and Environment, and Technology and Enterprise. Students may progress through the ordered outcomes at different rates within each learning area. The outcome approach was expected to improve student learning and provide a framework by which teachers and schools would be held accountable.

The Student Outcome Statements used in Western Australia grew out of a similar approach involving all Australian States and Territories and initiated by the Australian government in the early 1990s (Curriculum Corporation, 1994). Accountability was a central driving force with an integrated national approach. The Western Australian government wanted to produce its own outcome statements and did so using, in part, the national statements and profiles. Western Australian teachers trialed these in 1994/1995 (Education Department, 1996) and they were modified as a result of this trial. In 1997/1998, the Education Department has said that teachers may use the outcome statements approach on a voluntary basis and, after 2001, it will be mandatory for government schools as a system-wide change.

The Conceptual Framework of Teacher Receptivity

Teacher receptivity to the new system is assumed to be constructed from eight orientations (called variables in previous studies). In the present study, these orientations are evaluative attitudes towards the use of Student Outcome Statements, behavior intentions towards the use of Student Outcome Statements, feelings towards the use of Student Outcome Statements compared to the previous system, personal cost benefit of using Student Outcome Statements, support from significant others with teaching and resources, alleviation of concerns associated with the use of Student Outcome Statements, teacher collaboration in the use of Student Outcome Statements, and involvement in decision-making in the use of

Student Outcome Statements. These orientations should form a single scale measuring teacher receptivity to the use of Student Outcome Statements. Items of these orientations would be ordered on the scale from easy to difficult as they are all measuring some aspect of receptivity to the use of Student Outcome Statements.

The first six orientations are chosen because they have been shown to be related to teacher receptivity to other system-wide changes in Western Australia in previous studies of change (see Waugh and Collins, 1997; Waugh and Godfrey, 1995, 1993; Waugh and Punch, 1987, 1985). The last two orientations are new variables that have been shown to be important in major change overseas (see Hargreaves, 1994; Rosenholtz, 1991; Hargreaves, Davis, Fullan, Wignall, Stager and Macmillan, 1991). All these studies are correlational or qualitative in design. The present study uses a new approach to the analysis. It places all the items from the eight orientations of teacher receptivity on the same scale, provided that the items fit a Rasch (1960/1980) measurement model. The items fitting the model, (including attitudes, intentions and beliefs which are expected to be related, see Ajzen, 1989), will be ordered from easy to difficult (see Waugh and Collins, 1997). Using this measure of outcomes (teacher receptivity), it should then be possible to identify the important influences on receptivity which are not being addressed (the difficult items) and provide advice to administrators on how best to improve teacher receptivity to change.

The present study aims to create an interval level scale for the teacher receptivity to Student Outcome Statements, analyze its psychometric properties using a modern measurement model, the Extended Logistic Model of Rasch (Andrich, 1988a, 1988b; Rasch, 1960/1980), and provide advice to administrators on how best to improve teacher receptivity to system-wide changes.

Methods

Sample

Teachers from all Western Australian government secondary schools which were using Student Outcome Statements voluntarily in 1997 were invited to complete the questionnaire. The sample consists of 126 teachers: 66 (52%) from metropolitan schools and 60 (48%) from country schools. Of these, 115 (91%) were using Student Outcome Statements to plan teaching and learning programs at school and 100 (79%) were using Student Outcome

Statements for reporting student achievement to parents. As there are no official statistics for those teachers using the Student Outcome Statements, an unofficial estimate suggested that the response rate was about 60%.

Measurement Model

The Extended Logistic Model of Rasch (Andrich, 1988a, 1988b; Rasch, 1960/1980) is used with the computer program Rasch Unidimensional Measurement Models (RUMM) (Andrich, Sheridan and Luo, June, 1997) to analyze the data. This model unifies the Thurstone goal of item scaling with extended response categories for items measuring, for example, attitude, beliefs and feelings, which are applicable to this study. Item difficulties and person measures are placed on the same scale. The Rasch method produces scale-free person measures and sample-free item difficulties (Andrich, 1988b; Wright and Masters, 1982). That is, the differences between pairs of person measures and pairs of item difficulties are expected to be sample independent.

The RUMM program (1997) parameterizes an ordered threshold structure, corresponding with the ordered response categories of the items. The thresholds are boundaries located between the response categories and are related to the change in probability of responses occurring in the two categories separated by the threshold. A special feature of this version of the RUMM program is that the thresholds are re-parameterized to create an ordered set of parameters which are directly related to the Guttman principal components. With four categories, three item parameters are estimated: location or difficulty (δ), scale (θ), and skewness (η). The location specifies the average difficulty of the item on the measurement continuum. The scale specifies the average spread of the thresholds of an item on the measurement continuum. The scale defines the unit of measurement for the item and, ideally, all items constituting the measure should have the same scale value. The skewness specifies the degree of modality associated with the responses across the item categories.

The RUMM program substitutes the parameter estimates back into the model and examines the difference between the expected values predicted from the model and the observed values using two tests of fit: one is the item-trait interaction and the second is the item-person interaction.

The item-trait test of fit (a chi-square) examines the consistency of the item parameters across the person estimates for each item and data are combined across all items to give an overall test of fit. The latter shows the collective agreement for all items across persons of differing receptivity.

The item-person test of fit examines both the response pattern of persons across items and for items across persons. It examines the residual between the expected estimate and the actual values for each person-item summed over all items for each person and summed over all persons for each item. The fit statistics approximate a *t* distribution with a mean of zero and a standard deviation of one. Negative values indicate a response pattern that fits the model too closely (probably because dependencies are present) and positive values indicate a poor fit to the model (probably because 'noise' or other measures are present).

The Questionnaire and Data Collection

The first draft of the questionnaire was trialed with 15 secondary curriculum officers/teachers in the Education Department. The second draft was trialed with seven secondary school principals. Modifications in language, style and items were made resulting in 123 items for use as part of the original and wider data collection. Eighty-one items pertinent to the present study were taken for analysis: 68 items in a Likert format with four response categories and thirteen items in a semantic differential (see the appendix).

No neutral category was provided between agree and disagree because it would attract responses such as 'don't know', 'don't want to answer', 'unsure' and 'neutral', making interpretation unclear. However, a separate undecided category was provided for those who were genuinely undecided.

The eight orientations of teacher receptivity were measured by (see also the appendix):

1. Ten Likert style items on feelings towards the use of Student Outcome Statements compared to the previous system;
2. Five Likert style items on the personal non-monetary cost benefit of using Student Outcome Statements;
3. Eight Likert style items on perceived support for Student Outcome Statements by significant others;
4. Six Likert style items on behavior intentions towards the use of Student Outcome Statements;
5. Seven Likert style items on the alleviation of concerns with Student Outcome Statements;
6. Thirteen semantic differential style items on attitude towards Student Outcome Statements;

7. Eleven Likert style items on teacher collaboration; and
8. Ten Likert style items on teacher involvement in decision-making.

Data Analysis

The data were analyzed with the items for each of the eight orientations separately. Items which did not fit the model or had reversed thresholds (indicating that the categories were not answered consistently) were deleted. This was done because receptivity was designed and conceptualized from the eight orientations. Forty valid items were left from the original 81 items and these were analyzed together as one scale of teacher receptivity to Student Outcome Statements. All these 40 items fitted the model.

Results

The results are set out in one figure and six tables. Figure 1 shows the graph of teacher receptivity and item difficulties on the same scale in logits. Table 1 shows the location, scale and skewness values for 40 items fitting the model. Table 2 shows the threshold values of the 40 items fitting the model. Table 3 shows item-teacher interaction data. Table 4 shows the item-trait interaction data. Table 5 shows the location on the continuum, fit to the model and probability of fit to the model for the 40 items forming the scale, in location order. Table 6 shows teacher receptivity measures, standard errors and fit to the model in teacher measure order.

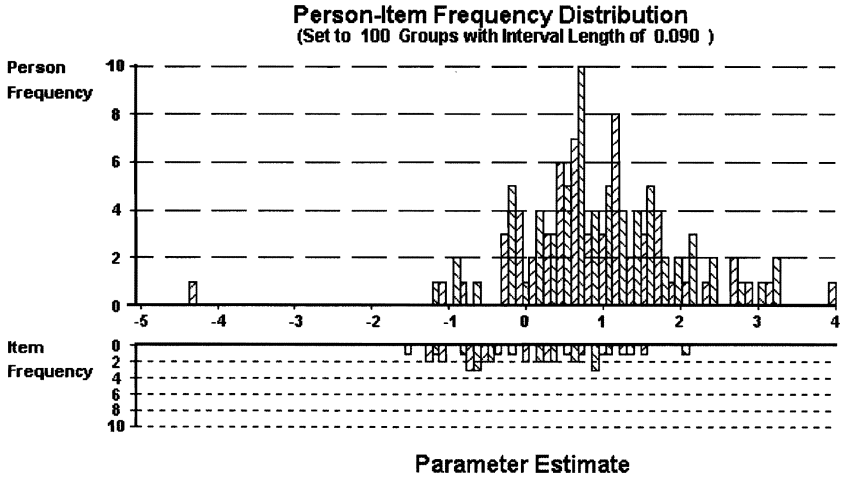


Figure 1. Graph of teacher receptivity and item difficulties.
Note: Teacher receptivity values and item difficulties can be read from the parameter estimate scale in logits.

Table 1

Location, Scale and Skewness values for 40 items fitting the model

	Location		Scale		Skewness	
Item Code	Estm	SE	Estm	SE	Estm	SE
Ex001 I001	-.701	0.175	1.009	0.109	.241	0.078
Ex002 I002	.618	0.161	1.441	0.160	-.059	0.054
Ex003 I003	-.657	0.152	.868	0.104	.027	0.072
Ex004 I004	-1.297	0.164	1.238	0.103	-.009	0.074
Ex005 I005	-.693	0.154	1.218	0.110	-.032	0.064
Ex006 I006	-.649	0.159	.750	0.103	.173	0.076
Ex007 I007	-.599	0.156	.821	0.104	.124	0.073
Ex008 I008	-.752	0.169	1.159	0.110	.115	0.072
Ex009 I009	.020	0.160	1.227	0.130	.028	0.061
Ex010 I010	-.412	0.146	.987	0.110	-.055	0.065
Ex012 I012	1.046	0.138	.937	0.122	.182	0.058
Ex013 I013	-.523	0.168	.960	0.109	.210	0.075
Ex022 I022	-.332	0.155	.764	0.105	.226	0.073
Ex024 I024	-.829	0.255	1.017	0.149	.264	0.117
Ex025 I025	.200	0.153	.723	0.130	-.138	0.074
Ex030 I030	-1.120	0.187	1.103	0.105	.353	0.086
Ex031 I031	.171	0.137	.697	0.106	.169	0.065
Ex032 I032	-1.253	0.177	1.324	0.105	.159	0.078
Ex035 I035	.881	0.121	.665	0.104	.013	0.058
Ex037 I037	.907	0.123	.597	0.106	.217	0.063
Ex040 I040	.680	0.147	.727	0.122	.159	0.068
Ex041 I041	.949	0.170	.842	0.147	.078	0.074
Ex054 I054	.669	0.115	.517	0.098	.094	0.060
Ex055 I055	-.012	0.135	.607	0.101	.195	0.067
Ex056 I056	.220	0.124	.546	0.100	.104	0.064
Ex057 I057	2.047	0.134	1.038	0.107	.119	0.058
Ex058 I058	1.338	0.121	.698	0.102	.035	0.058
Ex059 I059	1.517	0.121	.702	0.101	.084	0.059
Ex065 I065	-1.164	0.150	1.210	0.102	-.118	0.068
Ex067 I067	-.558	0.156	.792	0.102	.196	0.073
Ex068 I068	.513	0.146	1.017	0.124	.165	0.058
Ex069 I069	.330	0.149	.942	0.118	.223	0.062
Ex070 I070	1.277	0.136	1.021	0.119	.028	0.054
Ex071 I071	-1.123	0.188	1.315	0.110	.266	0.082
Ex072 I072	-1.545	0.176	.884	0.099	.203	0.086
Ex073 I073	-.186	0.148	.809	0.105	.212	0.067
Ex077 I077	-.485	0.142	.543	0.100	.117	0.074
Ex079 I079	.242	0.136	.814	0.108	.118	0.060
Ex080 I080	.911	0.114	.444	0.100	-.080	0.063
Ex081 I081	.354	0.132	.570	0.108	.128	0.067

No. of Items = 40 No. of Persons = 126
Item Convergence = 21 Person Convergence = 4
Convergence crt. = 0.010
Separation Index = 0.934

Psychometric Characteristics of the Receptivity Scale

The final 40 items measuring receptivity have a good fit to the measurement model, indicating a strong agreement between all 126 teachers to the different locations of the 40 items on the scale (see Table 5). That

Table 2

Threshold values of the 40 items fitting the model

	1	2	3
Ex001 I001	-1.536	-.965	2.502
Ex002 I002	-2.999	.235	2.764
Ex003 I003	-1.682	-.106	1.788
Ex004 I004	-2.495	.037	2.457
Ex005 I005	-2.501	.129	2.372
Ex006 I006	-1.155	-.692	1.847
Ex007 I007	-1.393	-.497	1.890
Ex008 I008	-2.087	-.462	2.548
Ex009 I009	-2.397	-.113	2.510
Ex010 I010	-2.085	.221	1.863
Ex012 I012	-1.509	-.728	2.237
Ex013 I013	-1.500	-.841	2.341
Ex022 I022	-1.075	-.905	1.980
Ex024 I024	-1.508	-1.055	2.562
Ex025 I025	-1.721	.550	1.171
Ex030 I030	-1.500	-1.410	2.910
Ex031 I031	-1.056	-.678	1.734
Ex032 I032	-2.329	-.636	2.965
Ex035 I035	-1.305	-.050	1.356
Ex037 I037	-.759	-.870	1.629
Ex040 I040	-1.136	-.636	1.772
Ex041 I041	-1.528	-.311	1.839
Ex054 I054	-.845	-.377	1.223
Ex055 I055	-.824	-.781	1.605
Ex056 I056	-.885	-.415	1.299
Ex057 I057	-1.838	-.478	2.316
Ex058 I058	-1.327	-.140	1.466
Ex059 I059	-1.236	-.337	1.573
Ex065 I065	-2.655	.472	2.183
Ex067 I067	-1.192	-.784	1.975
Ex068 I068	-1.706	-.658	2.364
Ex069 I069	-1.438	-.892	2.330
Ex070 I070	-1.986	-.112	2.097
Ex071 I071	-2.097	-1.066	3.162
Ex072 I072	-1.361	-.813	2.175
Ex073 I073	-1.194	-.847	2.041
Ex077 I077	-.852	-.469	1.321
Ex079 I079	-1.391	-.473	1.864
Ex080 I080	-1.048	.321	.727
Ex081 I081	-.884	-.513	1.397

Notes: The thresholds for each item are ordered from low to high corresponding to the ordering of the response categories. They are the unconstrained values.

is, there is strong agreement amongst the teachers to the item difficulties on the scale. The Index of Separability for the 40 item scale is 0.934 (see Table 4). The item threshold values are ordered from low to high indicating that the teachers have answered consistently with the ordered response format used (see Table 2). The item-trait tests of fit (see Table 5) indicate that the values of the item difficulties are strongly consistent across a

Table 3

Item-teacher interaction data

	Items		Persons	
	Location	Fit	Location	Fit
Mean	0.000	0.130	0.927	-0.338
SD	0.883	1.542	1.061	1.769
degrees of freedom = 119.88			degrees of freedom = 38.06	

Table 4

Item-trait interaction data

Total Item Chi Sq	130.341
Total Degree Freedom	78.000
Total ChiSq Probability	0.000
Person separation index	0.934
Test of Fit Power	EXCELLENT

range of teacher measures. The item-teacher tests of fit (see Table 4) indicate that there is good consistency of teacher and item response patterns. These data indicate that the errors are small and that the power of the tests of fit are good.

However, there is one problem area and this involves the scale values. The scale values for each item (observed average half threshold distance) vary too much (from 1.441 to 0.444 logits, see Table 1). In an ideal scale, these values should be equal, within the error measurement, as they define the unit of measurement. The variation probably arises from the measurement of 'noise'.

Meaning of the Receptivity Scale

The 40 items that make up the variable, teacher receptivity to Student Outcome Statements, are conceptualized from eight orientations. These eight orientations are confirmed as contributing to the variable. The 40 items define the variable. They have good content validity and they are derived from a conceptual framework based on previous research and theory. This, together with the previous data relating to reliability and fit to the measurement model, is strong evidence for the validity of the variable. This means that the teachers' responses to the 40 items are related sufficiently well to represent the variable teacher receptivity to Student Outcome Statements.

Table 5

Location on the continuum, fit to the model and probability of fit to the model for the 40 items forming the scale in location order

Label	Location	SE	Fit	ChiSq	Probability
Ex072 I072	-1.545	0.18	0.233	1.812	0.388
Ex004 I004	-1.297	0.16	-1.835	5.054	0.055
Ex032 I032	-1.253	0.18	-1.343	3.214	0.179
Ex065 I065	-1.164	0.15	1.463	2.453	0.274
Ex071 I071	-1.123	0.19	0.219	4.801	0.066
Ex030 I030	-1.120	0.19	-1.618	12.879	0.000
Ex024 I024	-0.829	0.26	0.293	0.128	0.936
Ex008 I008	-0.752	0.17	-0.962	2.719	0.237
Ex001 I001	-0.701	0.18	-1.025	6.088	0.022
Ex005 I005	-0.693	0.15	-1.023	3.376	0.163
Ex003 I003	-0.657	0.15	-1.155	5.800	0.030
Ex006 I006	-0.649	0.16	-2.369	9.395	0.000
Ex007 I007	-0.599	0.16	-2.178	5.872	0.028
Ex067 I067	-0.558	0.16	1.166	4.553	0.079
Ex013 I013	-0.523	0.17	-1.584	6.437	0.014
Ex077 I077	-0.485	0.14	-0.253	1.132	0.556
Ex010 I010	-0.412	0.15	-0.399	1.022	0.589
Ex022 I022	-0.332	0.16	-0.839	0.768	0.673
Ex073 I073	-0.186	0.15	-0.375	0.056	0.971
Ex055 I055	-0.012	0.14	-0.769	0.245	0.881
Ex009 I009	0.020	0.16	0.240	0.449	0.794
Ex031 I031	0.171	0.14	-0.816	0.312	0.852
Ex025 I025	0.200	0.15	1.473	0.409	0.810
Ex056 I056	0.220	0.12	-0.193	0.326	0.846
Ex079 I079	0.242	0.14	0.343	0.505	0.771
Ex069 I069	0.330	0.15	1.011	2.745	0.233
Ex081 I081	0.354	0.13	0.800	2.809	0.225
Ex068 I068	0.513	0.15	1.737	11.277	0.000
Ex002 I002	0.618	0.16	0.333	4.935	0.060
Ex054 I054	0.669	0.11	0.667	5.939	0.026
Ex040 I040	0.680	0.15	0.859	1.602	0.434
Ex035 I035	0.881	0.12	6.742	3.527	0.149
Ex037 I037	0.907	0.12	-0.143	0.494	0.775
Ex080 I080	0.911	0.11	0.778	1.951	0.360
Ex041 I041	0.949	0.17	0.138	3.944	0.116
Ex012 I012	1.046	0.14	0.328	0.229	0.889
Ex070 I070	1.277	0.14	2.104	6.069	0.023
Ex058 I058	1.338	0.12	1.209	2.198	0.315
Ex059 I059	1.517	0.12	0.777	1.521	0.453
Ex057 I057	2.047	0.13	1.164	1.294	0.511

Discussion of the Receptivity Scale

Figure 1 shows that most teachers find it easy to agree with most items; that is, the items are on the easy side and this means that teachers are receptive to the aspects of Student Outcome Statements represented by the items. Items at the easiest end of the scale (for example 72, 4, 32, 65, 71, 30, Table 5) are answered in agreement by nearly all the teachers.

Table 6

Teacher receptivity measures, standard errors and fit to the model in teacher measure order

Teacher no.	Total	Ability	SE	Fit	Teacher no.
31	2	-4.28	.761	1.758	31
88	29	-1.198	.283	.209	88
46	18	-1.056	.323	2.657	46
50	35	-.947	.245	4.112	50
56	39	-.893	.244	2.032	56
87	19	-.824	.385	-.039	87
47	48	-.613	.228	-.824	47
86	44	-.315	.242	-.887	86
3	53	-.297	.226	1.718	3
96	47	-.232	.242	-2.036	96
116	42	-.229	.252	-2.142	116
1	33	-.226	.269	.234	1
120	43	-.219	.25	-.656	120
53	58	-.141	.224	.207	53
107	54	-.141	.234	-1.37	107
124	55	-.083	.233	1.325	124
49	51	-.063	.246	-2.401	49
26	48	-.055	.257	1.109	26
109	35	-.051	.287	-1.8	109
48	44	-.005	.259	2.132	48
64	48	.087	.262	.527	64
114	31	.098	.293	-.189	114
106	58	.144	.24	.885	106
28	58	.151	.239	-1.51	28
25	62	.164	.227	-.382	25
79	57	.193	.244	.024	79
99	54	.241	.264	-.563	99
121	66	.264	.226	-.875	121
110	53	.285	.251	-1.243	110
71	57	.328	.247	-1.868	71
42	50	.331	.27	.102	42
102	59	.333	.245	-.91	102
97	64	.41	.237	1.152	97
68	57	.422	.249	-.398	68
23	61	.427	.246	.765	23
100	66	.438	.234	-1.117	100
123	39	.46	.297	-4.239	123
83	70	.471	.228	.083	83
67	66	.498	.241	-3.29	67
126	68	.515	.237	-2.313	126
98	53	.517	.272	-1.801	98
39	66	.527	.239	.251	39
55	58	.54	.259	-.312	55
4	71	.607	.234	1.267	4
75	59	.621	.271	-5.7	75
105	73	.629	.231	-.391	105
84	68	.637	.24	-.314	84
62	71	.638	.232	.978	62
51	64	.648	.246	-3.327	51
54	63	.656	.261	-3.592	54
72	70	.676	.237	-4.067	72
60	74	.682	.231	.852	60
30	74	.682	.231	-.661	30
40	65	.696	.249	-4.308	40
45	66	.71	.252	-3.921	45
94	70	.712	.236	.14	94
10	71	.722	.239	-.929	10
104	72	.728	.24	-.251	104
37	70	.758	.243	-2.575	37
91	57	.76	.279	-2.334	91
63	66	.79	.249	-.811	63
22	66	.837	.256	-.034	22

(*Table continues*)

(*Table 6 continues*)

Teacher no.	Total	Ability	SE	Fit	Teacher no.
29	77	.845	.234	-.899	29
5	75	.862	.237	1.511	5
21	47	.869	.286	-1.616	21
85	76	.913	.237	1.341	85
16	69	.923	.251	-2.216	16
27	79	.955	.236	-1.111	27
118	79	.955	.236	-4.023	118
2	72	1.016	.248	-1.415	2
69	75	1.041	.253	1.162	69
34	74	1.046	.247	.734	34
70	81	1.068	.238	3.377	70
125	81	1.068	.238	1.682	125
89	69	1.093	.271	-1.953	89
122	82	1.125	.24	-1.364	122
113	82	1.125	.24	-2.21	113
76	80	1.143	.242	2.13	76
35	70	1.157	.262	-.882	35
11	81	1.164	.245	-3.286	11
74	76	1.174	.251	1.768	74
81	74	1.182	.252	-.804	81
8	83	1.183	.241	-.587	8
80	73	1.217	.256	.108	80
13	73	1.247	.269	.071	13
33	78	1.25	.254	-1.001	33
32	80	1.255	.252	3.013	32
66	76	1.365	.268	-.999	66
111	78	1.367	.259	-2.042	111
65	80	1.414	.258	-2.303	65
19	64	1.437	.279	-.19	19
101	83	1.459	.256	1.182	101
82	77	1.478	.269	-2.259	82
103	82	1.504	.26	1.563	103
6	86	1.522	.253	-1.91	6
78	65	1.547	.305	-.46	78
108	87	1.573	.255	-.08	108
38	81	1.594	.268	.542	38
43	81	1.608	.269	.094	43
115	88	1.608	.259	1.244	115
44	81	1.617	.268	3.931	44
52	91	1.675	.256	3.183	52
95	72	1.681	.285	1.844	95
9	61	1.722	.305	2.309	9
73	87	1.73	.265	-1.925	73
90	88	1.765	.268	-.911	90
24	86	1.778	.273	-1.053	24
41	87	1.89	.276	.073	41
93	95	1.947	.267	1.322	93
119	96	2.019	.27	-.113	119
7	97	2.093	.274	-.212	7
77	86	2.158	.297	1.563	77
92	98	2.169	.277	-1.051	92
12	96	2.183	.283	.342	12
57	95	2.317	.293	.118	57
112	82	2.393	.333	.684	112
17	101	2.411	.291	.335	17
61	91	2.679	.332	1.63	61
20	104	2.68	.309	1.194	20
59	100	2.793	.326	.033	59
36	101	2.893	.333	-.031	36
14	105	3.073	.344	.016	14
117	108	3.102	.342	-1.06	117
58	109	3.223	.354	.134	58
15	104	3.258	.368	-.988	15
18	111	3.992	.454	.035	18
Person Mean:	0.927	Fit Mean:	-0.338		
Person SD :	1.061	Fit SD :	1.769		

This means, for example, that teachers found it easy to agree that they participated in selecting resources for Student Outcome Statements (item 72, Table 5); that Student Outcome Statements address the needs of individual students better than in the previous system (item 4); that in their behavior and communication with others, they intend to say that Student Outcome Statements are useful for planning teaching and learning programs (item 32); that other teachers seek their advice about teaching problems (item 65); and that teachers share ideas with teachers not in their department at school (item 71). The outcomes here are that teachers find it easy to agree with these items. They are contributing positively to teacher receptivity and teachers are receptive to these aspects related to Student Outcome Statements. Similar comments can be made about the other easy items. It should be noted that these results also mean that administrators have done a good job explaining and preparing for those aspects relating to the easy items. Teachers are supportive of these and they have contributed to their positive receptivity.

Items at the hard end of the scale (for example item nos. 57, 59, 58, 70, 12, 41 Table 5) are only answered in agreement by those teachers who have high receptivity (for example teacher nos. 18, 15, 58, 117, 14, 36, Table 6). Teachers whose receptivity is at the low end of the scale (for example nos. 31, 88, 46, 50, 56, 87, Table 6) would not be able to answer in agreement with the hard items. This means, for example, that teachers found it hard to agree that the Student Outcome Statements are uncomplicated (item 57); that they are time efficient (item 59); that they are clear (item 58); that they offer advice about teaching to other teachers, without being asked for it (item 70); that, in weighing up the balance between the extra work generated by Student Outcome Statements and homelife, Student Outcome Statements are worthwhile (item 12); and that they can access the District Office to obtain advice about Student Outcome Statements (item 41). Similar comments can be made about other hard items.

The information from this scale leads directly to advice that can be given to administrators and decision-makers about how to improve teacher receptivity (attitudes and intentions) to Student Outcome Statements. There are four difficult items relating directly to the alleviation of concerns about Student Outcome Statements which imply that improvements could be made and which, in turn, could lead to improvements in teacher receptivity. One, administrators can make it easier for teachers to have access to Central Office in regards to contact, information and support. Two, principals can hold school meetings at which teachers can raise their concerns about Student Outcome Statements and have those concerns dealt with quickly and

efficiently at school. Three, principals and senior teachers need to organize for someone at school to provide support when teachers have implementation problems with Student Outcome Statements. Four, the District Superintendent can organize for teachers to have easy contact and information support when problems with Student Outcome Statements arise.

There are four aspects of Student Outcome Statements that can be improved which, in turn, could lead to improved teacher receptivity. Administrators could make the system less complicated, more time efficient for teachers, more clear and more realistic. If this is difficult to do, it may be possible for administrators and principals to explain the system to teachers again.

There are six other aspects which are moderately difficult (items 2, 68, 81, 69, 79, 25) and which could be addressed to improve receptivity. Administrators could arrange professional development to show teachers how they could manage their classrooms better with the new system. Principals could encourage teachers to share advice on teaching problems and help each other more with problems relating to the new system at school. Principals could help organize professional development for teachers that focus on problems and issues directly related to the new system, as needed by teachers. School administrators can allow and help teachers to be a little more flexible in implementing the system to suit the needs of students with difficulties.

Conclusion

The Extended Logistic Model of Rasch was useful in creating a scale of teacher receptivity to Student Outcome Statements (a system-wide educational change in a centralized system) and for investigating the psychometric properties of the scale. The evidence suggests that the scale has excellent reliability and validity. The analysis confirms the conceptual design of teacher receptivity as involving at least eight orientations: evaluative attitudes, behavior intentions, feelings towards the change compared to the previous system, the benefits of the change, support from significant others, alleviation of concerns, collaboration with other teachers, and involvement in decision-making.

The analysis leads to the conclusion that teacher receptivity to the change can be improved by focusing on the difficult items which represent the important influences not addressed by administrators. Hence administrators can make strong improvements to teacher receptivity to this change

by improving teachers' access to information from Central Office, by holding school meetings to allow teachers to raise specific concerns, and by providing help to teachers with implementation problems. Moderate improvements to receptivity can be made through the use of professional development relating to the change, encouraging teachers to share advice on problems and solutions, and for administrators to be more flexible in allowing teachers to suit the needs of students with difficulties. Receptivity could also be improved if administrators could make the change less complicated, easier to implement and more realistic for classroom use.

Acknowledgments

Special thanks are given to Rose Moroz, Superintendent of Education, Cannington District, Perth, for permission to use her data on teacher receptivity to Student Outcome Statements.

References

Ajzen, A. (1989). Attitude, structure and behaviour. In A, Pratkanis, A, Breckler and A, Greenwald (Eds.), *Attitude, structure and function.* New Jersey: Lawrence Erlbaum and Assoc.

Andrich, D., Sheridan, B. and Luo, G. (1997). RUMM: a windows-based item analysis program employing Rasch Unidimensional Measurement Models. Perth: Murdoch University, School of Education.

Andrich, D. (1988a). *Rasch Models for Measurement.*

Andrich, D. (1988b). A General Form of Rasch's Extended Logistic Model for Partial Credit Scoring. *Applied Measurement in Education, 1*(4), 363-378.

Curriculum Corporation (1994). National statements and profiles. Melbourne: Curriculum Corporation.

Education Department of Western Australia (1994). Student Outcome Statements (working edition). Perth: Education Department

Education Department of Western Australia (1996). Report of the Student Outcome Statements trial 1994-1995 (executive summary. Perth: Education Department

Hargreaves, A. (1994). *Changing teachers, changing times.* London: Cassell

Hargreaves, A., Davis, T., Fullan, M., Wignall, R., Stager, M. and Macmillan, R. (1991). Secondary work cultures and educational change. Ontario: OISE

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

Rosenholtz, S. (1991). *Teachers' workplace: the social organization of schools.*

New Yok: Longman

Waugh, R.F. and Collins, P. (1997). Catholic school teachers' receptivity to a proposal to move Year 7 primary classes to secondary schools. *Education Research and Perspectives*, 24 (1), pp 63-83

Waugh, R. F. and Godfrey, J. (1995). Understanding teachers' receptivity to system-wide change. *Journal of Educational Administration, 33* (3), pp 39-56

Waugh, R. F. (1994). Teachers' receptivity to system-wide change in a centralised education system. *Education Research and Perspectives, 21* (2), pp 80-94.

Waugh, R. and Godfrey, J. (1994). Measuring students' perceptions of cheating. *Education Research and Perspectives, 21* (2), pp 28-37.

Waugh, R. F and Godfrey, J. (1993). Teacher receptivity to system-wide change in the Implementation stage. *British Education Research Journal*, 19 (5), 565-578.

Waugh, R.F. and Punch, K.F. (1987). Teacher receptivity to system wide change in the implementation stage. *Review of Educational Research, 57*(3), 237-254.

Waugh, R.F. and Punch, K.F (1985). Teacher receptivity to system wide change. *British Education Research Journal, 11*(2), 113-121.

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

Appendix

40 Items of the Questionnaire fitting the model

In comparison to the previous system (Unit Curriculum), the use of Student Outcome Statements allows me to:

1. provide for better student learning.
2. manage my classroom better.
3. provide more relevant content.
4. address the needs of individual students better.
5. provide more varied experiences for the students.
6. better describe student learning.
7. make better judgements about student achievements.
8. plan more relevant learning experiences for my students.
9. demonstrate my accountability.
10. report more effectively on student achievement.

Non-monetary cost benefit of using Student Outcome Statements.

12. In weighing up the balance between any extra work generated for you by Student Outcome Statements and your home life, the use of use of Student Outcome Statements is worthwhile.
13. In weighing up the balance between any extra work generated for you by Student Outcome Statements and better student classroom learning, the use of Student Outcome Statements is worthwhile.

Significant other support for using Student Outcome Statements.

22. Most teachers in this department support Student Outcome Statements.
24. The District Superintendent supports Student Outcome Statements.
25. Most teachers in this school support Student Outcome Statements.

In my behavior and communication with others, I will probably say that Student Outcome Statements are useful:

30. for monitoring student achievement.
31. for reporting student achievement to parents.
32. for planning teaching and learning programmes.

Alleviation of concerns about the use of Student Outcome Statements.

35. There are regular school meetings at which I can raise my concerns about Student Outcome Statements.
37. There is good general school support whenever I have problems with the implementation of Student Outcome Statements in the classroom.
40. I can access Central Office support to obtain advice about Student Outcome Statements.
41. I can access District Office support to obtain advice about Student Outcome Statements.

Attitudes towards the use of Student Outcome Statements

54.	realistic	idealistic
55.	effective	ineffective
56.	necessary	unnecessary
57.	uncomplicated	complicated
58.	clear	unclear
59.	time efficient	time inefficient

Teacher collaboration

65. Teachers seek my advice about their teaching problems in this department.
67. I share teaching resources and materials with teachers who are not in my department.
68. Teachers who are not in my department seek my advice about their teaching problems.
69. If I have a teaching problem, I get advice from teachers who are not in my department.
70. I don't offer advice to teachers about their teaching unless I am asked for it.
71. I share ideas with teachers who are not in my department.

Involvement in decision-making

72. Teachers participate in selecting instructional materials and resources in this department.
73. Teachers participate in determining the content of professional development sessions in this department.
77. I am involved in decisions that are related to Student Outcome Statements in this department.
79. Teachers participate in determining the type of of whole school professional development that we have in this school.
80. I am involved in decisions outside my department which are related to Student Outcome Statements.
81. Teachers are encouraged by a deputy-principal to modify the curriculum to meet student needs.

JOURNAL OF OUTCOME MEASUREMENT®, 3(1), 89–102
Copyright© 1999, Rehabilitation Foundation, Inc.

Distractors—Can They Be Biased Too?

Sivakumar Alagumalai

John P. Keeves

The Flinders University of South Australia

Numerous work has been done on item bias and differential item functioning. Although there is some research on distractor analysis, no detailed study has been attempted to examine the way distractors in an item function, with regards to comparing distractor performance. This paper examines how distractors function differentially and compares various methods for identifying this. The Pearson chi-square, likelihood ratio chi-square and Neyman weighted least squares chi-square tests are some of these methods. Possible causes of distractor bias are discussed with illustrations from a physics problem-solving scale.

Requests for reprints should be sent to Sivakumar Alagumalai, School of Education Stut Campus, The Flinders University of South Australia, Bedford Park, SA5042, Australia. e-mail: sivakumar.alagumalai@flinders.edu.au

Introduction

Items in a test have been known to be biased against particular sub-groups and have become a cause of concern to examinees, users of tests and the testing community (Hambleton and Swaminathan, 1985; Thissen, Steinberg and Fitzpatrick, 1989). Thorndike (1982) argues that bias is involved whenever the group with which a test is used brings to the test a cultural background noticeably different from that of the group for which the test was primarily developed and standardized.

Although bias is referred to differently by different researchers, Osterlind (1983, p. 4) defines it as:

> a systematic error in the measurement process. It affects all measurement in the same way changing measurement – sometimes increasing it and other times decreasing it ... Bias than is nothing more or less than consistent distortion of statistics.

In current work involving item response theory (IRT), the terms 'differential item functioning' (DIF) and 'differential item performance' (DIP) are used instead of item bias. However, it must be noted that DIF/DIP is not restricted to IRT approaches (Fan, 1998; Rudner, Getson and Knight, 1980a).

Although the term DIF (DIP) indicates that items function differently for various subgroups of examinees, and can be argued to be different from bias (Scheuneman and Slaughter, 1991), the terms are used interchangeably in this paper as supported by Johanson and Johanson (1996). Furthermore, the aim of the paper is to highlight the normally forgotten micro-level bias of the distractors as compared to the much discussed and formulated item-level bias. The paper also seeks to review associated statistics in order to identify distractor level bias and is illustrated with items selected from a physics problem-solving test administered in two countries.

Distractor Bias—the Roots

Numerous techniques have been advanced for identifying bias at the item level. Matuszek and Oakland (1972) compared the factor structures for subgroups taking the same items. Green and Draper (1972) used the item-score point biserials for subgroups and identified the biased items. However, Wright, Mead and Draba (1976, p. 4) indicate the limitations of these methods and argue that, "any technique which relies upon correla-

tions as the indication of bias is vulnerable to the variation on the trait in the samples studied. Where groups differ in their traits variability, then items can appear biased when they are not." Rudner, Getson and Knight (1980) investigated seven biased item detection techniques. Their study examined two transformed item difficulties approaches, three item characteristic curve (ICC) approaches, and two chi-square approaches. They found that the "chi-square technique was found to be effective as the three parameter ICC theory technique" (Rudner, Getson and Knight, 1980a, p. 4)

Adams (1984) has discussed three methods of identifying bias. In the *ANOVA* method, the focus is on the interaction between group membership and correct response. The item difficulty is examined and the measure of bias is the significance of F, both its main effects and interactions.

The focus of the transformed item difficulties method is parallel to the ANOVA method. However, the measure of bias is some arbitrary designation of distance on a scatter plot. In addition to these two methods there is the chi-square method and its focus of analysis is on the difference in proportion attaining a correct response across total score categories. The significance of chi-square is the measure of bias and is based on item difficulty. Although these methods are useful, the estimation procedures are based on classical test theory and on the proportion attaining a correct response (Adams, 1992).

Contrary to classical test theory, IRT uses the assumption that examinee performance on a test can be predicted from a defined examinee characteristic, by estimating scores for the examinee on the characteristic and using these scores to predict or explain item and test performance (Lord, 1980). The relationship between examinee item performance and the set of traits assumed to be influencing the characteristic and item performance can be described by a monotonic item characteristic function (ICF). In the one-parameter and single dimensional model, the ICF is common to all items and is called item-characteristic curve (ICC). It provides the probability of examinees answering an item correctly at different points on the ability scale (Hambleton and Swaminathan, 1985).

Lord and Stocking (1997) highlight the advantage IRT has to provide a natural method for detecting item bias as the item response function, in theory, does not depend on the group used for calibration. In order to detect item bias, the item response functions of the target subgroups are compared. Items that were biased would have curves that were significantly different (Lord, 1980). Adams (1992) argues that the difference

or area between these curves for the two subgroups would be better for determining item bias, as the area is an estimate of probability of success of examinees of equal ability level.

Apart from comparing the area under the ICC, numerous other indices for estimating item bias have been reported (Thissen, Steinberg and Fitzpatrick, 1989). The item difficulty, also referred to as the p-value of an item, is the measure of proportion of examinees in a given population or sub-population who answer correctly (Osterlind, 1983).

Another approach in IRT uses the discrimination index, or the item fit statistics to assess the degree to which an item correctly differentiates between examinees on a test. The discrimination index is used to indicate the power of an item in separating the capable from the less capable on a specified latent attribute.

Numerous studies (Bezruczko, et al., 1989; Wang and Lane, 1994) have utilized these indices for identifying item bias. The *compare* routine in the QUEST (Adams and Khoo, 1993) program produces the difference of item difficulties ($d_1 - d_2$), the standardized item difficulties ($std\_d_1 - std\_d_2$), the associated chi-square values and the corresponding p-values. Biased items are commonly considered to be those whose (i) difference in difficulty is below –0.50 or greater than +0.50; (ii) difference in adjusted standardized level of difficulty indices between the foci groups fall out of the range ±2.00; and (iii) discrimination indices of the groups under survey are below 0.15 or greater than 0.55. However, of these three criteria, the second is heavily dependent on sample size, and the third involves a reversion to the use of classical test theory, in the context of using item response theory. In all the studies cited above, once an item has been identified as biased, discussion is primarily based on the generalities of the item, especially with the item-stem. Very little emphasis has been given to the alternatives themselves, and how they may have caused bias or functioned differentially. It can thus be argued that the methods and indices discussed thus far restrictively look into the proportion of items correct rather than considering all the alternative responses wholistically. Rudner, Getson and Knight (1980a) indicate that item bias is distinct from the issue of test bias. Hence, it can be argued along similar lines that distractor bias warrants special attention as it is different from item bias. The next section outlines possible methods for identifying distractor bias, and highlights conditions where an item may be biased while the alternatives are not and vice-versa.

Theoretical Consideration, Bayes' Theorem and Associated Goodness-of-Fit Indices

An important application of Bayes' theorem is in the analysis of case-control studies (Freeman, 1987). This study design gathers data on gender type (cases) and compares the choice of alternatives found in a comparable or control group. From categorical analysis of data theory, it is possible to compare the conditional probabilities of gender and to assess the relationships between gender and the selection of alternatives which are considered as categories here. Freeman (1987) argues that what is required in such situations is the comparison of the conditional probabilities of alternative selection, subject to the conditional probabilities for gender. Use of Bayes theorem, moreover, permits the latter design, also called the cohort study, to be linked to the former case-control study design, and the incidence rates can be reported as conditional probabilities.

From these conditional probabilities of incidence rates, Freeman (1987, p.38) shows that a probability distribution can be hypothesized, regardless of whether the parameters are estimated or specified. Thus the data obtained lend themselves to the use of tests of goodness-of-fit. The Pearson chi-square test (χ^2) automatically allows for comparison of parameters. Freeman (1987) further argues that the likelihood ratio chi-square (G^2) and the Neyman weighted least squares chi-square (Q) are also useful for assessing the fit for data-specified distributions. The following forms were used in computing distractor bias in this paper:

$$\chi^2 = \Sigma^{I}_{i=1}((Y_i - m_i)^2 / m_i) \quad (1)$$

$$G^2 = 2\,\Sigma^{I}_{i=1}(Y \ln (Y_i / m_i)) \quad (2)$$

$$Q = \Sigma^{I}_{i=1}((Y_i - m_i)^2 / Y_i) \quad (3)$$

where Y_i is the observed count

$m_i = E(Y_i / H_0)$ [m_i is the expected count for Y in the *i*th category when the hypothesised distribution is true (H_0), $i = 1, \ldots I$.]

Tabachnick and Fidell (1996, p.245) caution against the use of χ^2 as its nonadditivity becomes a serious issue as additional variables produce higher order associations. They suggest that an alternative strategy is to use the likelihood ratio statistic G^2, which has the property of additivity of effects. Further to this, Cohen et al. (1996) have indicated that the G^2 could be used to detect DIFs and the findings would be consistent with the one-parameter Rasch model. The Q-statistic has been found to be use-

ful in categorical data analysis and Freeman (1987) argues that all three goodness-of-fit statistics should be reported both for comparability reasons and for consistency checks.

It must also be recorded here that the goodness-of-fit tests used in this study are different from the chi-square procedure used by Adams (1992). He argues that chi-square is particularly sensitive to within-groups item discriminations, as well as being constrained by the arbitrary selection of ability levels, ability being estimated from only those who responded correctly to an item (Adams, 1992, p.181). The goodness-of-fit procedures used here are based on data-specified distributions and take into consideration all the alternatives as categories. Scheuneman (1979) also argues that the chi-square procedures are rough approximations to the IRT model. Hence, the chi-square statistical procedures referred to above should enable the meaningful interpretation of differential functioning of distractors, as a "concerted method" of examining bias (Marascuilo and Slaughter, 1981, p. 229).

The Year 10 Physics Problem Solving Test

In the years 1995 and 1996, approximately 650 Year 10 (Secondary Four) students from South Australian and Singaporean schools were tested in a 25 multiple-choice item test in physics problem solving. The data were analyzed with the Rasch procedure and all items fitted the model (INFIT MNSQ range $0.83 \leftrightarrow 1.20$). The data were then subjected to the *compare* routine available in the QUEST (Adams and Khoo, 1993) program which identified the following items as biased:

Item No 6: (*easier for males*)

One half-second after starting from rest, a freely falling body will have a velocity of about

A.	2.5 m s^{-1}	C.	10 m s^{-1}
B.*	5 m s^{-1}	D.	20 m s^{-1}

Items 13 & 14 refer to the following information:

A ball is thrown vertically upwards with an initial velocity of 10 m s^{-1}. Neglect air resistance.

Item No 13: (***easier for females***)

What is the maximum height reached by the ball?

A.	2 m	C.	10 m
B.*	5 m	D.	100 m

Item No 14: (***easier for males***)

How long is the ball in the air?

A.	1 s	C.	5 s
B.*	2 s	D.	10 s

Table 1 summarizes the relevant statistics produced through the *compare* routine.

The findings in Table 1 are consistent with the goodness-of-fit indices computed with equations (1), (2) and (3) and are summarized below.

Table 1

QUEST (Adams and Khoo, 1993) output for compare *routine*

Item #	Adjusted Delta males d1	Adjusted Delta females d2	Difference d1-d2	Difference d1-d2 std'ised	Chi-Sq	p
6	-0.94	-0.01	-0.94	-5.11	26.12	0.00
13	0.88	0.46	0.41	2.36	5.57	0.02
14	0.63	1.25	-0.62	-3.45	11.89	0.00

The χ^2, Q and G^2 values for Items 6, 13 and 14 are significant ($p<0.05$) for the associated degrees of freedom (df=3). Although the differential functioning of distractors could be examined through Table 2, distractors-ability plots (D-A Plots) were drawn to aid in this task. Figure 1 and 2 are the distractors-ability plots for Items 6, and 13 respectively.

For Item 6 it is evident from the plots that there are no interactions between the distractors and sex of students. There is a consistent pattern in the distribution of mean performance or ability levels across the distractors. Each distractor has about an equal amount of pull of students in a particular performance or ability level and is reflected in a parallel manner on the axis of the other subgroup. The significant values of the goodness-of-fit statistics could partly be attributed to relatively more females (31%) were drawn to alternative C. Forsyth and Spratt (1980) contend that reading skills, process skills and computational skills are three

Table 2

Response rates and goodness-of-fit indices for Items 6, 13 and 14

Alternatives		A	B	C	D	df	χ^2	Q	G^2
Item #6	M %	4.1	74.5*	15.0	5.1	3			
	M_Ab	-0.61	0.51	-0.33	0.17		**26.44**	**29.53**	**27.00**
	F_%	5.9	57.5*	31.1	5.6	3			
	F_Ab	-0.48	0.67	-0.03	0.39				
Item #13	M %	1.3	36.9*	51.6	9.2	3			
	M_Ab	0.12	0.92	0.01	-0.30		**11.72**	**12.59**	**11.86**
	F_%	3.2	47.5*	40.2	8.8	3			
	F_Ab	-0.64	0.88	0.01	-0.36				
Item #14	M %	43.6	42.0*	4.1	8.9	3			
	M_Ab	0.10	0.72	-0.04	-0.27		**9.31**	**9.54**	**9.34**
	F_%	52.2	31.7*	6.7	8.8	3			
	F_Ab	0.32	0.72	-0.33	-0.07				

Note: M_%: Males–Percentage
M_Ab: Mean Males–Ability Level in Logits
F_%: Females–Percentage
F_Ab: Mean Females–Ability Level in Logits

cognitive skills that differentiate between individuals with respect to problem-solving ability. Thus, it is possible that females tended to merely recall (process skill) the concept that acceleration due to gravity is 10 m s^{-2} and simply choose alternative C even though the units were not equivalent, where as males choose alternative B (the correct option) which seemed more viable than alternative C as only 15 per cent of males chose it.

Item 13, however, shows an interaction effect with alternative A being chosen by on average higher performing males (0.12 logits) compared to their female counterparts (-0.64 logits). Girls of performance or ability level 0.01 logit tended to choose alternative C. Thus for alternative A alone there exists a difference between males and females in average performance level of about 0.74 logits. The differential choice rate towards this wrong alternative may be due in part to the conceptual proximity of option A (2 metres) which seemed more realistic to males. However, females of about equal performance or ability level thought otherwise. Option A was chosen by the relatively weaker females (-0.64).

The findings here are in contrast to a study done by Lawrence and Curley (1989) on the response of males and females to items pertaining to

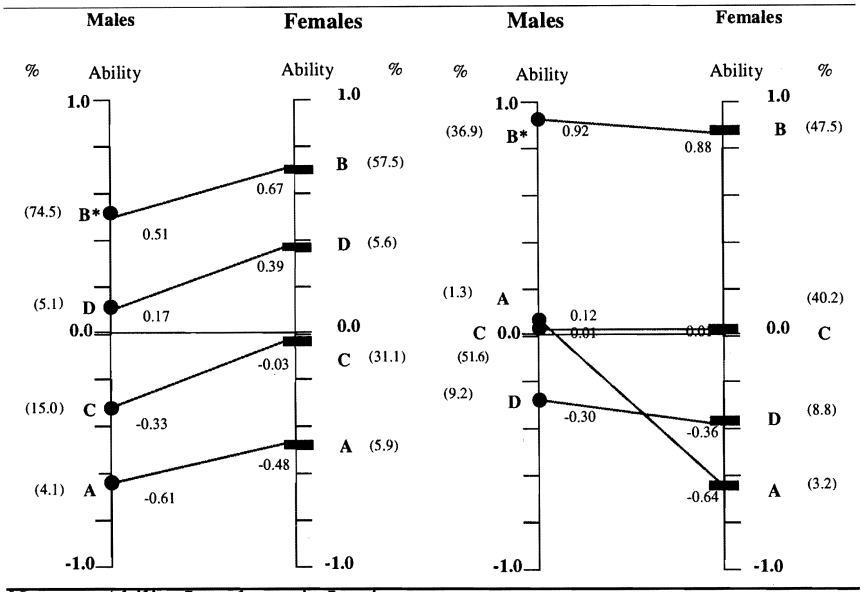


Figure 1: D–A Plots for Item 6

Figure 2: D–A Plots for Item 13

Note: Ability Levels are in Logits
A, B, C and D are alternatives

technical aspects of science. They found that items that involved technical information were more difficult for females. However, in the case of Item 13, females found the item relatively easy. In addition, the distractors were attracting the sexes to respond differently.

It is evident from these two sample items that even though an item may be identified as biased through IRT procedure, the goodness-of-fit statistics employed coupled with the D-A plots helps to identify further whether bias also occurs at the distractor level.

The hypothesis that distractors may be behaving differentially rather than the item itself was extended beyond the items identified as biased by the *compare* routine in QUEST (Adams and Khoo, 1993). Two items in the physics problem-solving test exhibited characteristics similar to Items 6 and 13 although not identified as biased items by the program. Table 3 summarizes the output for the *compare* routine.

The chi-square values for the comparison between males and females are not significantly different and this implies that Items 11 and 18

Table 3

QUEST (Adams & Khoo, 1993) output for ***compare*** *routine*

	Adjusted Delta		Difference			
Item #	males d1	Females d2	d1-d2	d1-d2 std'ised	Chi-Sq	p
11	0.03	-0.25	0.27	1.57	2.46	0.12
18	0.41	0.16	0.25	1.43	2.03	0.15

are not biased toward any particular subgroup. However, an examination using the analysis of distractors produced different results and are summarized in Table 4.

Although the items were not biased (as bias was not identified by the IRT software), the goodness-of-fit statistics show that significant differences exist between the two subgroups. In order to identify the functioning of the distractors, D-A plots were done for Items 11 and 18 and are given in Figure 3 and 4 respectively.

In order to facilitate understanding and interpretations of the information presented above, details of Items 11 and 18 are provided.

Item 11 refers to the following information:

A rope is tied to a mass of 8.0 kg which is at the bottom of a 10 m cliff. A person pulls the mass up to the top of the cliff at a constant speed of 2.0 m s^{-1}.

Table 4

Response rates and goodness-of-fit indices for Items 11 and 18

Alternatives		A	B	C	D	df	χ^2	Q	G^2
Item #11	M %	15.6	55.1*	23.2	4.1	3			
	M_Ab	-0.29	0.72	-0.02	-0.62		**11.08**	**11.80**	**11.17**
	F_%	8.2	62.5*	21.1	6.5	3			
	F_Ab	-0.34	0.69	-0.03	-0.30				
Item #18	M %	16.6	46.8*	18.8	13.4	3			
	M_Ab	-0.10	0.78	-0.01	-0.01		**9.68**	**10.20**	**9.72**
	F_%	9.1	54.0*	22.0	12.6	3			
	F_Ab	-0.36	0.82	-0.03	-0.21				

Note: M_%: Males-Percentage
M_Ab: Mean Males-Ability Level in Logits
F_%: Females-Percentage
F_Ab: Mean Females-Ability Level in Logits

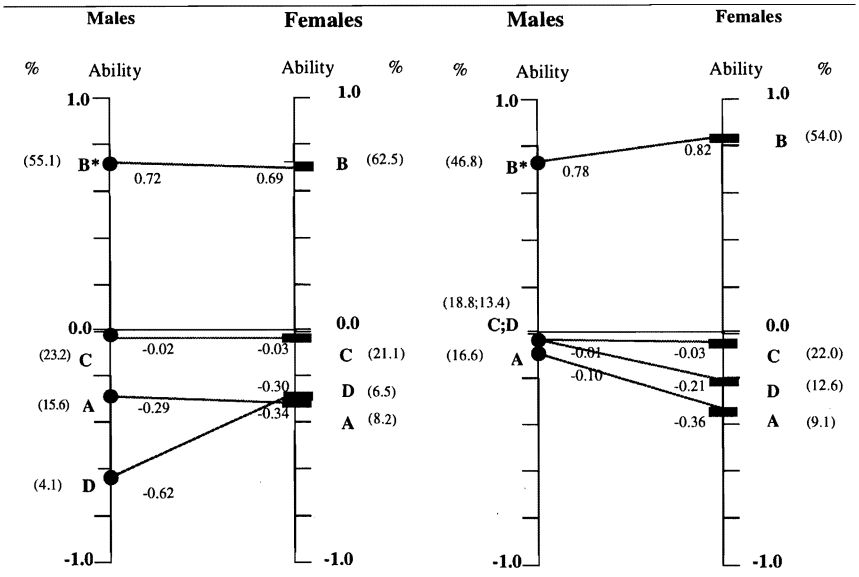


Figure 3: D–A Plots for Item 11

Figure 4: D–A Plots for Item 18

Note: Ability Levels are in Logits
A, B, C and D are alternatives

Item No 11: (***neutral***)

At what average power does the person work?

A.	40 W	C.	400 W
B.*	160 W	D.	1600 W

Item No 18: (***neutral***)

A pole AB of length 10 m and weight 800 N has its centre of gravity 4m from the end A, and lies on horizontal ground. The end B is to be lifted by a vertical force applied at B. What is the least force required to do this?

A.	200 N	C.	640 N
B.*	320 N	D.	3200 N

With reference to Figure 3 and Item 11, females on average of nearly equal ability (-0.30 and –0.34) selected options D and A, and this contrasted with the ability levels of males who chose the same distractors. Thus options A and D pose a greater challenge to females than to males, as a 0.33 logit difference exist between males who chose option A as compared to those who chose D. Thus, distractors A and D are function-

ing differentially for the two subgroups under study and may have contributed significantly to the difference between the subgroups, as illustrated by high values in goodness-of-fit statistics employed.

The reverse seems to be true for Item 18 shown in Figure 4, equal ability male students (-0.01) are caught between choosing alternative C or D, where as the female students who opted for these two alternatives had about 0.18 logit difference between them. In both Items 11 and 18, it was noted that both males and females had failed to notice the clue given through the units provided beside the numeric solution. Furthermore, the approximate distance between the best and the other groups varied between the male and female subgroups.

The above technique highlights the advantage of examining distractor bias through the chi-squared approach as it is based on the distribution of correct responses across ability levels. Rudner, Getson and Knight (1980b), have argued that this technique is not restricted to identifying bias for only one entire group but also for all associated subgroups.

Discussion and Conclusion

Although the underlying reasons for why particular distractors attract a particular subgroup more than the other can be debated at length, the above methods of using χ^2, Q and G^2 simultaneously highlight bias at the distractor level that may have been missed by conventional IRT software. The chi-square methods employed above is relatively easier to implement as compared to the methods advocated by researchers examining DIF. Furthermore, extrapolation of the χ^2, Q and G^2 statistics for distractors is facilitated and simplified through the use of EXCEL spreadsheets.

Furthermore, the use of distractors-ability level plots for subgroups graphically represents the choice of alternatives by the groups and highlights probable operating responses of the groups under examination. The distractors-ability plot is independent of indices of DIF/DIP and in addition opens up the possibility of calculating 'shift-indices' that may be indicative of true differential functioning of distractors, as the ability values are calculated independently of the sample and the items employed and thus reflect strong and meaningful estimates of the performance of the subgroups.

The use of categorical data analysis techniques complements the findings based on the above plots and allows for consistent comparison between statistical procedures. Above all, this paper highlights the need

for all those involved with tests and testing to attempt to look at distractor-level differential functioning prior to devoting considerable thought to the item- and test-level biases as a test with substantial distractor bias indicates that different subgroups are responding differently. If this is so, then the test scores cannot be interpreted in the same manner for the different subgroups. However, the biggest challenge still exits in the reporting of these biases and/or differential functions at both school and class-levels.

References

Adams, R.J. (1984). *Sex Bias in ASAT?* ACER Research Monograph No. 24. ACER, Hawthorn, Victoria.

Adams, R.J. (1992). *Item Bias.* In J.P. Keeves (Ed.) *The IEA Technical Handbook.* The International Association for the Evaluation of Educational Achievement (IEA), The Hague, The Netherlands.

Adams, R.J. and Khoo, S.K. (1993). *QUEST – The Interactive Test Analysis System.* ACER, Hawthorn, Victoria.

Bezruczko, N., et al., (1989). The stability of four methods for estimating item bias. *Paper presented at the Annual Meeting of the American Educational Research Association.*

Cohen, A.S., et al., (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement,* 20(1), 15-26.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381.

Forsyth, R.A. and Spratt, K.F. (1980). Measuring problem-solving ability in mathematics with multiple-choice items: The effects of item format on selected items and test characteristics. *Journal of Educational Measurement, 17*(1), 31-43.

Freeman, D.H. (1987). *Applied Categorical Data Analysis.* Marcel and Dekker Inc, New York.

Green, D.R. and Draper, J.F. (1972). Exploratory studies of bias in achievement tests. *Paper presented at the American Psychological Association Annual Convention.* Honolulu, Hawaii, September 1972.

Hambleton, R.K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Kluwer Nijhoff Publishing, Boston, MA.

Johanson, G.A. and Johanson, S.N. (1996). Differential Item Functioning in Survey Research. *Paper presented at the Annual Meeting of the American Educational Research Association.* New York, N.Y., April 8-12, 1996.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Erlbaum Lawrence, Hillsdale, N.J.

Lord, F.M., and Stocking, M.L. (1997). *Item Response Theory.* In J.P. Keeves (Ed.) *Educational Research, Methodology and Measurement: An International Handbook.* Pergamon Press, Oxford.

Marascuilo, L.A. and Slaughter, R.E. (1981). Statistical procedures for identifying possible sources of item bias based on c^2 statistics. *Journal of Educational Measurement, 18*(4), pp.229-248.

Matuszek, P. and Oakland, T. (1972). A factor analysis of several reading readiness measures for different social, economic and ethnic groups. *Paper presented at the American Educational Research Association*, Chicago, Illinois, 1972.

Osterlind, S.J. (1983). *Test Item Bias.* Sage University Paper Series on Quantitative Applications in the Social Sciences. SAGE Publication, London.

Rudner, L.M., Getson, P.R. and Knight, D.L. (1980a). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Education Measurement, 17*(1), pp. 1-10.

Rudner, L.M., Getson, P.R., and Knight, D.L. (1980b). Biased item detection techniques. *Journal of Educational Statistics, 5*(3), 213-233.

Scheuneman, J.D. (1979). A method for assessing bias test items. *Journal of Education Measurement, 16*(2), 143-152.

Scheuneman, J.D. and Slaughter, C. (1991). *Issues of Test Bias, Item Bias, and Group Differences and what to do while waiting for the answer.* ERIC Document: ED 400294

Tabachnick, B.G. and Fidell, L.S. (1996). *Using Multivariate Statistics.* Harper Collins College Publishers, New York.

Thissen, D., Steinberg, L., and Fitzpatrick, A.R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26*(2), 161-176.

Thorndike, R.L. (1982). *Applied Psychometrics.* Houghton-Mifflin, Boston, MA.

Wang, N. and Lane, S. (1994). Detection of gender-based differential item functioning in a mathematics performance assessment. *Paper presented at the Annual Meeting of the National Council on the Measurement in Education.* New Orleans, L.A., April 3-7, 1994.

Wright, B.D., Mead, R. and Draba, R. (1976). *Detecting and Correcting Test Item Bias with a Logistic Response Model.* MESA Psychometric Laboratory Research Memorandum Number 22, University of Chicago.

CONTRIBUTOR INFORMATION

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

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

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

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

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

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

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

SUBSCRIBER INFORMATION

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

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