JOURNAL OF OUTCOME MEASUREMENT®

Volume 2, Number 2 1998

JOURNAL OF OUTCOME MEASUREMENT®, 2(2), 79 - 96

The Functional Assessment Measure (FAM) in Closed Traumatic Brain Injury Outpatients: A Rasch-Based Psychometric Study

Luigi Tesio
Salvatore Maugeri Foundation, IRCCS, Dept. of Research, Functional Assessment and Quality Assurance in Neuromotor Rehabilitation, and Dept. of Rehabilitation, Pavia, Italy

Anna Cantagallo
Unità Operativa di Medicina Riabilitativa, Arcispedale Sant'Anna, Azienda Ospedaliera, Ferrara, Italy

The Functional Assessment Measure (FAM) has been proposed as a measure of disability in post-acute Traumatic Brain Injury (TBI) outpatients. It is comprised of the 18 items of the Functional Independence Measure (FIM℠), scored in terms of dependence, and of 12 newly designed items, scored in terms of dependence (7 items) or performance (5 items). The FIM℠ covers the domains of self-care, sphincter management, mobility, locomotion, communication and social cognition. The 12 new items explore the domains of community integration, emotional status, orientation, attention, reading/writing skills, swallowing and speech intelligibility. By addressing a set of problems quite specific for TBI outpatients the FAM was intended to raise the ceiling of the FIM℠ and to allow a more precise estimate of their disability. These claims, however, were never supported in previous studies. We administered the FAM to 60 TBI outpatient, 2-88 months (median 16) from trauma. Rasch analysis (rating scale model) was adopted to test the psychometric properties of the scale. The FAM was reliable (Rasch item and person reliability 0.91 and 0.93, respectively). Two of the 12 FAM-specific items were severely misfitting with the general construct, and were deleted. Within the 28-item refined FAM scale, 4 new items and 2 FIM℠ items still retained signs of misfit. The FAM was on average too easy. The most difficult item (a new one, Employability) did not attain the average ability of the subjects. Also, it was only slightly more difficult than the most difficult FIM℠ item (Memory). The FAM does not seem to improve the FIM℠ as a far as TBI outpatients are to be assessed.

Requests for reprints should be sent to Luigi Tesio, Fondazione Maugeri, Divisione di Recupero e Rieducazione Funzionale, Via A.Ferrata 8, 27100 Pavia, Italy.

The Functional Assessment Measure (FAM) (Hall, 1992; Hall, Hamilton, Gordon, & Zasler, 1993; Hall, Mann, High, Wright, Kreutzer, & Wood, 1996) is a 30-item 7-level ordinal scale proposed as a measure of disability in post-acute traumatic brain injury (TBI) patients. It is composed by the 18 items of the Functional Independence Measure (FIM^{SM}) to which 12 newly designed items were added (Table 1). The FIM^{SM} is an international standard instrument (Granger, Hamilton, Linacre, Heinemann, & Wright, 1993; Tsuji, Sonoda, Domen, Saitoh, Liu, & Chino, 1995) for measuring disability. This is defined as "any restriction or lack....of ability to perform an activity in the manner or within the range considered normal for a human being" (WHO, 1980), and thus reflects the overall capacity of interaction of the whole person with his/her environment (Tesio, 1997). The FIM^{SM} covers the sub-domains of personal care, sphincter control, transfer mobility, locomotion, communication and social cognition. The scale is anchored by extreme ratings of total dependence (level 1) and total independence (level 7) and considers amount of another person's assistance (levels 1 to 4), supervision (level 5), and use of adaptive devices (level 6). The FIM^{SM} was designed to measure the burden of care caused by disability, regardless of the underlying disease. It was designed to be targeted on the level of disability affecting post-acute rehabilitation inpatients. There are sound demonstrations of its reliability and validity. In particular, Rasch-based rating scale analysis (Linacre, Heinemann, Wright, Granger, & Hamilton, 1994) investigated its unidimensionality and showed that the global fit of the items to a unique continuum is satisfactory, but it is greatly improved by analyzing two subsets of items distinctly: a motor subscale (items 1 to 13) and a cognitive subscale (items 14 to 18). The proponents' expectance was also confirmed, that the hierarchy of item difficulty would be highly consistent across different impairments, including brain injury (Granger, Hamilton, Linacre, Heinemann, & Wright, 1993; Heinemann, Linacre, Wright, Hamilton, & Granger, 1993; Linacre, Heinemann, Wright, Granger, & Hamilton, 1994).

The FAM was designed to better capture the specificity of disability of TBI patients, mainly in the post-discharge phase. The first concern was to raise the "ceiling" of the FIM^{SM}. FIM^{SM} scores, in fact, saturate when patient approaches independence in basic daily activities, which should be the case for most patients at discharge. A second concern was to increase the FIM^{SM} sensitivity and precision in TBI outpatients by addressing some motor and cognitive problems specific for this condition. Twelve more items were thus designed (Table 1). Like for the FIM^{SM}, seven of

Table 1
The Functional Assessment Scale (FAM)

Functional Independence Measure FIM℠	Newly-designed items
Eating	Swallowing
Grooming	Car transfer
Bathing	Community access
Dressing upper body	Reading*
Dressing lower body	Writing*
Toileting	Speech intelligibility*
Bladder management	Emotional status*
Bowel management	Adjustment to limitation*
Bed, chair, wheelchair transfer	Employability
Toilet transfer	Orientation
Tub, shower transfer	Attention span*
Walking, wheelchair	Safety judgement
Stairs	
Comprehension	
Expression	
Social interaction	
Problem solving	
Memory	

Note. The 30-item Functional Assessment Measure (FAM) instrument is made-up by combining the 18-item Functional Independence Measure (FIM℠, left) with the 12 newly designed items (right). Twenty-five out of the 30 items are scored from 1 to 7 with 1 being the lowest number of dependence and 7 the highest number of independence. The 5 items marked with an asterisk are scored higher the more physiologic the patient's performance.

these items were anchored by extreme ratings of total dependence and independence (levels 1 through 7). For 5 out of 12 items, however, grading reflected increasing performance, not independence.

The FIM[SM] has been adopted in countless published studies, aimed both at investigating its psychometric properties (Linacre, Heinemann, Wright, Granger, & Hamilton 1994; Ottenbacher, Yungwen, Granger, & Fiedler, 1996) and at measuring disabilities in the most various impairments (Granger, Hamilton, Linacre, Heinemann, & Wright, 1993; Heinemann, Linacre, Wright, Hamilton, & Granger, 1993). Measures were also used for epidemiologic and economics studies (Stineman, Escarce, Goin, Hamilton, Granger, & Williams, 1994; Tesio, Perucca, Franchignoni, & Porta, 1996). By contrast, only a few published studies addressed the validity of the FAM (Hall, Hamilton, Gordon, & Zasler, 1993; Hall, Mann, High, Wright, Kreutzer, & Wood, 1996). The former research showed that the FAM has satisfactory interrater reliability.

In the same study a Rasch-analysis on TBI patients was cited. Details were not provided. The Authors simply reported that "the FAM items were more widely spread than FIM[SM] items." The most difficult items appeared to be Community transfer, Stairs and Car transfer. It was also reported that "there is a redundancy in measuring the disability, both within the FIM[SM] and for the FIM[SM]+FAM items." The second study evidenced that both the FIM[SM], the FAM and a well known questionnaire of community integration (the Community Integration Questionnaire-CIQ) all suffer from a marked ceiling effect when applied to brain injury outpatients. There are reasons, therefore, to suspect that the FAM instrument adds nothing to its FIM[SM] component. This study specifically addresses this hypothesis.

METHODS

Patients

One author (AC), specialist in Physical Medicine and Rehabilitation and in Neurology, visited and tested all of the patients. Their clinical picture is summarized in Table 2. We enrolled 60 consecutive outpatients (40 men, age 16-65, median 27) attending the physiatric referral of a large teaching hospital in Ferrara, Italy. All patients had suffered from closed Traumatic Brain Injury (TBI) for which they had been admitted to an intensive care and/or neurosurgery unity. Afterwards, they had been transferred to a dedicated post-acute rehabilitation unit. The initial condition could be clas-

Table 2
Demographic and Clinical Description of 60 TBI Outpatients

TBI Group (40 M; 20 F)	Mean	SD	Median	Range
Age (years)	30.7	13.3	27	16-65
Education (years)	10.5	3.8	8	3-17
GCS (3-15)	5.5	1.7		3-9
Coma Duration (days)	40.9	48.4	23.5	1-280
PTA* Duration (weeks)	18.1	8	25	1-25
TBI - FAM Interval (months)	23.7	21.3	16	2-88
FIM Score (18-126)	104	19.2	107.5	37-126
FAM Score (30-210)	168.7	30.1	173.5	75-209

*Post Traumatic Amnesia (PTA)

Impairments	/60	%
Motor	47	78
Sensorial	29	48
Cognitive:		
Attention	19	32
Memory	31	52
Language	25	42
Problem-Solving	11	18
More than one Cognitive Impairment	34	57
Reentry to Work	15	25

sified as severe or moderate (Glasgow Coma Scale , GCS, < 7 or 8-12) in 42 and 5 subjects, respectively. In the remaining 13 patients the initial GCS score was not available. In these cases the coma lasted for more than one day, supporting a "severe"more than a "moderate" classification. Median duration of coma had been 23.5 days, range 1-280. Subjects came at the outpatient referral after 2-88 months, median 16. During the outpatient visit, the presence of any motor and/or sensory impairment was recorded. Cognitive impairments were recorded through standardized neuropsychological tests (Spinnler & Tognoni, 1987). Deficits in long-term memory, language fluency, attention and problem-solving were detected in 31, 25, 19 and 11 cases, respectively. Fifteen, only, out of the 60 patients had returned to work. Injuries to other body regions, frequent in TBI patients, had left meaningful impairments in only one subject. He presented with weakness of the upper limbs following peripheral nerve lesions (see also Results).

Measures

Instruments: All of the subjects were administered the FAM instruments through direct interview by one author (AC). Family members or significant proxies were also interviewed, whenever indicated and possible. The Italian validated version of the FIM^{SM} was adopted (Uniform Data System for Medical Rehabilitation, 1993), following the official guidelines. The rater had achieved a competency certificate after attending a one-day course held in Italy under license from the Copyright owner agency (UB Foundation, State University of New York, Buffalo, NY). The Manual of the two FAM-specific items was followed in its original American form (Hall, 1992).

Analysis: the Rasch model: The response matrix was subjected to Rasch analysis, rating scale model, through the BIGSTEPS software (Linacre, & Wright, 1993). The Rasch technique is an item-response model, allowing to measure both the item difficulty and the subject's ability along a shared continuum (Andiel, 1995). The model has a unique set of properties, making it an ideal tool for testing the validity of ordinal scales (Wright, & Linacre, 1989).

a) Difficulty and ability can be estimated independently from each other, along a shared continuum of "less" to "more".

b) Measures are provided in true equal-interval units which is usually expressed in logit form (Wright & Stone, 1979). Conventionally a zero-

measure represents the average difficulty of the items. Items sharing the same measures are suspect to be redundant, in that thy indicate the same level of "disability" (in the FAM case), although they represent different performances. A "physical" analogy might be provided by a test requiring the subject to jump over a 50 cm-high chair, then over a 50 cm-high wooden box and over a 50 cm-high stone. Using three different items, rather than one, does not make our estimation of jumping capacity more precise.

c) The model also corrects the unavoidable "compression" of ordinal scores in the vicinities of the scale extremes.

d) The model is probabilistic, not deterministic. To each subject/item interaction a given probability is ascribed. This allows to estimate the precision of the computed measure of difficulty/ability.

e) The model is prescriptive, not descriptive. The need for coherence of the individual items with the measure's underlying construct (unidimensionality) is emphasized.

The theory assumes that only the interaction of subject's ability and item difficulty along a unique variable should give rise to a reasonable matrix of responses. The probability of response is then modelled. On the observed responses, the matrix most compatible with the model is estimated. For each item and each subject the "fit" of the actual string of responses to the model-expected pattern is calculated. "Misfit" occurs when an item accumulates unlikely responses across subjects or, on the subject's side, when a person accumulates unlikely responses across items. Misfit can be considered either too large (positive) or too small (negative), with respect to the response variance expected by the model (Wright, & Linacre, 1994; Wright & Masters, 1982; Wright, & Stone, 1979). Positive misfit comes from the accumulation of unexpected responses (e.g. when a given items is passed by less able subjects, and/or it is not passed by more able subjects). Negative misfit comes from too predictable patterns, lacking the expected variance: e.g. when a given item is always passed by more able subjects, and it is never passed by less able subjects. This statistic is most sensitive to unexpected responses when item and subject measures are very different (e.g. when a very easy item is missed by a very able subject). Therefore, it is also called "outfit." Anomalous subjects are the most common source of large misfit indices. Guessing and idiosyncracies (i.e. individual, usually unknown reasons for passing or failing a given item) are common sources of large positive and negative outfit indices, respectively. Misfit is more difficult to detect when the item difficulty is "on target," and thus most informative, on the subject's ability (i.e. when item

and subject have comparable measures). In fact, failure and success are both expected with high probability. In these cases, an "information weighted" misfit statistic, called "infit," is most sensitive to unexpected patterns of response. Anomalous items are the most common source of large infit indices. An ambiguous and/or generic definition can make the item representative of one or more extraneous domains. This can make the pass/fail transition within an item response string either too erratic (large positive infit) or too sharp (large negative infit) if an extraneous domain captures some peculiarities of the subjects' sample.

f) The model can be coherently applied to dichotomous as well as to polychotomous items. In these latter, measure and fit of the within-item steps can also be investigated either by assuming that the same hierarchy of step measures be shared by the various items (rating scale model) or that an item-specific step hierarchy exists (partial credit model) (Wright & Masters, 1982; Wright, & Stone, 1979). The levels of within-item steps can thus be treated like the items, with respect to the definition of their measure and fit statistics. Removing and/or collapsing redundant or misfitting levels may often lead to an improvement of the metric properties of the overall scale.

g) The BIGSTEPS software also provides a "person separation index" and an "item separation index" defined as the ratio of true spread of the measures with their measurement error. A clinically useful scale should encompass at least 3 strata of patients (e.g. severe, moderate, mild, or so), implying an index greater than 2.0. On the item side, the greater the separation, the greater is range of disability (in the case of the FAM) which can be measured by the scale. A related index is the reliability of these separation indices (range 0-1) which is defined as the ratio of true score variance divided by (error+true) variance. Coming from the model, this statistics is sample and rater independent, and does not require actual multiple testing procedures, like conventional reliability indices (e.g. Cohen's k or ICCs).

Investigation of the scale structure: This study was focused on the scale properties, not on subjects' features. According to the literature, we defined as acceptable for both items and persons (studies on small samples) outfit or infit mean square indices (MNSQ) >1.4 or < 0.6, and/or standardized indices (ZSTD) >2 or < -2 (Wright, & Linacre, 1994). The Rasch analysis then followed its typical trial-and-error diagnostic procedure, aiming at diagnosing the cause for misfit (Wright, & Masters, 1982; Wright, & Stone, 1979). Misfitting subjects and/or misfitting FAM specific (not FIM^{SM}) items were interactively eliminated. The step structure of the FAM was also

investigated, in order to optimize the overall scale fit (Wright & Masters, 1982). For example, under utilized steps and/or steps representing higher measures than suggested by their definition (e.g. a level 3 representing a higher measure than a level 4) are usually worth to be collapsed ore rescored. It was decided not to delete any misfitting items of the FIM[SM] within the FAM. In fact, the aim of the 12 FAM-specific items is extending the range of measures covered by the FIM[SM], taken as a reference standard.

RESULTS

Figure 1 gives the frequency distribution of the FAM raw scores. It is quite evident that the FAM suffers from a marked ceiling effect: it is too easy with respect to the population analyzed.

Table 3 gives a BIGSTEPS output page. This shows the best possible overall mix of item infit and outfit statistics we could obtain. Level 1 and 2 in the FAM items were misfitting, and the average difficulty of step 2 was lower, compared to step 1. These two steps could be hardly related to different ability levels, and were thus collapsed. This gave rise to an orderly increasing hierarchy of difficulty among the remaining 6 levels. We also had to remove 17 misfitting subjects out of the original 60 (see below). The 30 FAM items are listed, from top to bottom, in order of decreasing measure (more difficult items on top, easier items on bottom). The 12 FAM-specific items are underlined. Four out of the 12 FAM-specific items show marked signs of misfit which are: *Community Access, Adjustment to limitations, Emotional status, and Writing.* The most troublesome items appear to be *Emotional status and Adjusting to limitations*, which have both high positive infit and outfit statistics. The same happens to 2 out of the 18 FIM[SM] items, i.e. *Memory and Social Interaction.* Item and person measure reliability were satisfactory (see legend).

Table 4 gives the best possible overall item fit we could obtain after removing Adjusting to limitations and Emotional status. One more subject, getting an extreme score, was deleted. Two of the 10 residual FAM-specific items (Community Access, Writing) and two FIM[SM] items (Memory, Social Interaction) still show high infit and outfit statistics. The item difficulties span over 3 logits, usually a clinically useful range. There are signs of redundancy in the –1/+1 difficulty range. It may be seen that different items (i.e. Eating and Toilet Transfer, as well as, Orientation, Car Transfer and Grooming) share the same measures. Item and person reliability measures are satisfactory (see legend).

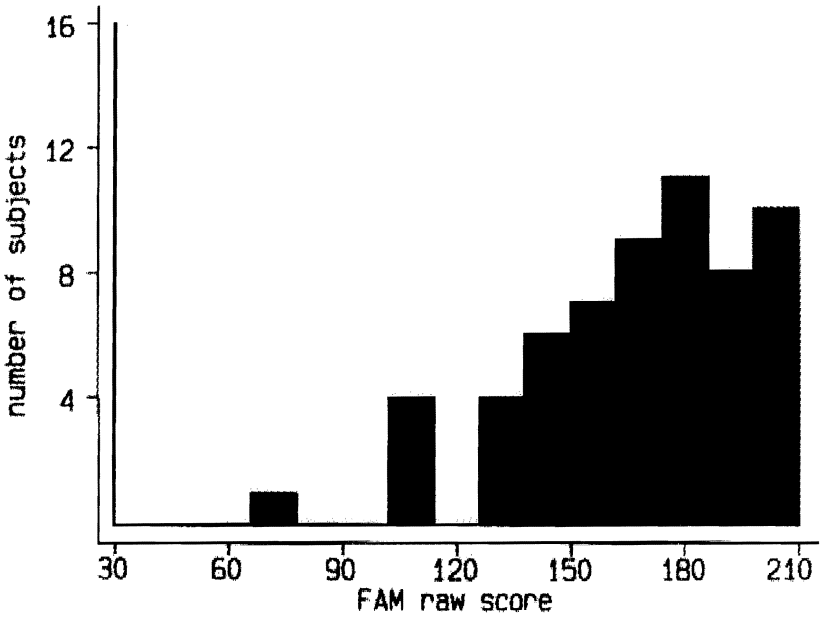


FIGURE 1. Frequency distribution of FAM scores across 60 TBI outpatients.

Table 3
Rasch Analysis of FAM Scale

			INFIT		OUTFIT			
	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	PTBIS	ITEMS
	1.50	.14	.54	-2.7	.71	-1.4	.76	Employ
	1.10	.14	1.31	1.4	1.39	1.5	.66	CommAcc
	1.00	.14	1.63	2.5	1.51	1.9	.59	Mem
	.79	.14	1.54	2.2	1.53	1.8	.55	AdjLim
	.77	.14	1.41	1.7	1.93	3.0	.49	Emotion
	.47	.15	1.28	1.2	1.62	2.0	.56	Write
	.45	.15	1.02	.1	.83	-.7	.78	DressLow
	.45	.15	1.00	.0	1.20	.7	.64	T/ShTr
	.38	.15	.98	-.1	.90	-.4	.75	Bath
	.38	.15	1.05	.2	.98	-.1	.64	Attn
	.33	.15	1.18	.7	1.10	.4	.64	SafJudg
	.21	.16	.85	-.7	1.14	.5	.55	PrSolv
	.08	.16	.92	-.3	.75	-.9	.79	Toil
	.06	.16	.47	-2.8	.47	-2.2	.74	Read
	.03	.16	.62	-1.8	.53	-1.9	.83	DressUp
	-.08	.17	1.09	.3	.86	-.5	.70	Stairs
	-.11	.17	.97	-.1	.76	-.8	.71	Groom
	-.11	.17	.88	-.5	.68	-1.1	.79	CarTransf
	-.11	.17	.94	-.2	.86	-.5	.64	Orient
	-.20	.17	1.70	2.3	1.85	2.1	.39	SocInt
	-.32	.18	.83	-.7	.66	-1.1	.76	Walk/Wch
	-.39	.18	1.11	.4	1.04	.1	.48	SpeechInt
	-.42	.19	.80	-.8	.77	-.7	.53	Expr
	-.53	.19	.58	-1.9	.44	-1.9	.80	ToilTr
	-.68	.20	.66	-1.4	.75	-.7	.61	Eat
	-.68	.20	.82	-.7	.57	-1.3	.74	B/C/Wch
	-.72	.20	1.22	.7	.88	-.3	.63	Blad
	-1.16	.24	1.95	2.4	1.03	.1	.62	Bowel
	-1.21	.24	.90	-.3	.95	-.1	.43	Compr
	-1.28	.25	1.25	.7	.94	-.1	.42	Swallow
MEAN	.00	.17	1.05	.1	.99	-.1		
S.D.	.67	.03	.35	1.4	.38	1.3		

Note. The Rasch analysis of the FAM scale is applied to 60 TBI outpatients. The table is taken, slightly simplified, form the output pages of BIGSTEPS software (MESA Press). Seventeen misfitting patients were removed from the original sample. The items are listed from bottom to top in order of increasing difficulty in the right column. (For full item names, refer to Table 1.) The FAM-specific items are underlined. From left to right the following variables are also listed: Item Measure, Standard Error, "Infit" and "Outfit" statistics (both mean-square and standardized), and Point Biserial Correlation Coefficient. The two bottom rows give Mean and Standard Deviation of the values recorded in the corresponding columns. Rasch item real separation coefficient 3.32, reliability 0.92. Rasch person separation coefficient 3.64, reliability 0.93 (Wright, & Linacre, 1994; Wright, & Masters, 1982).

Table 4
Rasch Analysis of FAM Scale
With Misfitting Items and Misfitting Patient Removed

			INFIT		OUTFIT			
	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	PTBIS	ITEMS
	1.63	.15	.67	-1.7	.88	-.5	.72	Employ
	1.21	.15	1.36	1.5	1.56	2.0	.64	CommAcc
	1.10	.15	1.75	2.9	1.69	2.4	.55	Mem
	.55	.15	1.32	1.3	1.58	1.9	.55	Write
	.52	.15	1.03	.1	.85	-.6	.78	DressLow
	.52	.15	.99	.0	1.16	.6	.65	T/ShTr
	.45	.16	.93	-.3	.84	-.6	.77	Bath
	.45	.16	1.17	.7	1.11	.4	.60	Attn
	.40	.16	1.23	.9	1.16	.5	.63	SafJudg
	.28	.16	.97	-.1	1.29	1.0	.50	PrSolv
	.14	.17	.91	-.4	.75	-.9	.80	Toil
	.12	.17	.51	-2.5	.55	-1.8	.72	Read
	.09	.17	.59	-2.0	.51	-2.0	.85	DressUp
	-.03	.17	1.06	.2	.85	-.5	.72	Stairs
	-.06	.17	.92	-.3	.73	-.9	.74	Groom
	-.06	.17	.90	-.4	.70	-1.1	.79	CarTransf
	-.06	.17	1.01	.0	.94	-.2	.62	Orient
	-.15	.18	1.87	2.7	2.37	3.1	.34	SocInt
	-.28	.18	.83	-.7	.66	-1.1	.77	Walk/Wch
	-.35	.19	1.16	.6	1.10	.3	.48	SpeechInt
	-.38	.19	.88	-.5	.88	-.4	.49	Expr
	-.49	.20	.56	-1.9	.44	-1.9	.81	ToilTr
	-.65	.21	.66	-1.4	.76	-.7	.63	Eat
	-.65	.21	.83	-.7	.59	-1.3	.75	B/C/Wch
	-.70	.21	1.23	.8	.94	-.1	.63	Blad
	-1.15	.24	1.97	2.5	1.06	.1	.64	Bowel
	-1.21	.25	.97	-.1	1.06	.1	.40	Compr
	-1.27	.25	1.34	1.0	1.04	.1	.39	Swallow
MEAN	.00	.18	1.06	.1	1.00	-.1		
S.D.	.69	.03	.36	1.3	.41	1.2		

Note. This is the same analysis shown in Table 3 with two misfitting items (Emotional Status and Adjusting to Limitations) and one misfitting patient removed. Rasch item real separation coefficient 3.27, reliability 0.91. Rasch person separation coefficient 3.59, reliability 0.93.

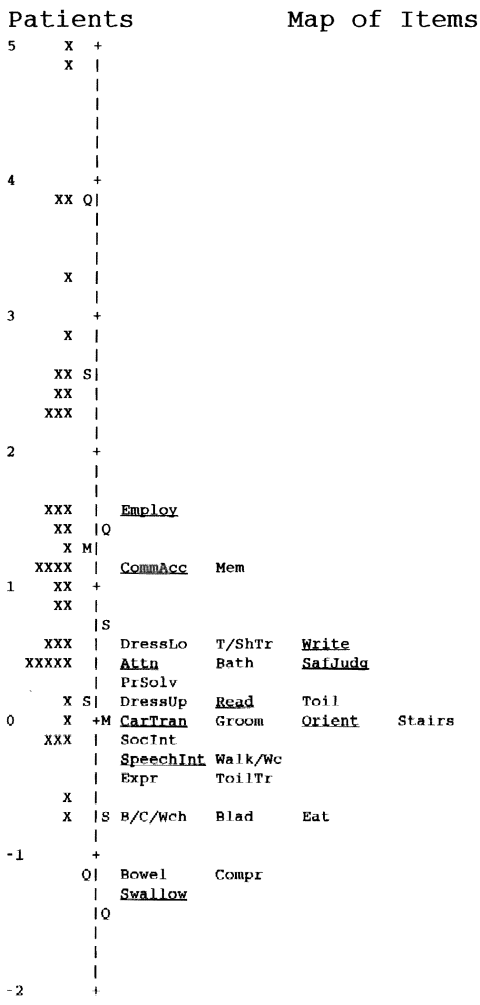


FIGURE 2 "Item map" provided by Rasch analysis through BIGSTEPS software. Forty-three TBI outpatients ("X" symbols: 42 fitting patients and one patinet with extreme, maximal score) and the 28 items of refined FAM scale (2 misfittings items deleted, see Table 4) are aligned along a shared measurement continuum of "disability." Measurement units (true interval logit units) are given on the left. A higher measure means higher item difficulty or greater subject's ability. "M," 'S," and :Q" symbols indicate mean, 1SD and 2SD of the measures recorded in patients and items, respectively. The underlined items are FAM-specific, in that they were added to the original 18 items of the FIM.

Figure 2 gives the BIGSTEPS "maps of items" of the refined 28-item refined FAM. Items (on the right) and subjects (on the left) are aligned along the measurement continuum. Higher measures (logit units, on the left) represent greater subject's ability or item difficulty. It's quite evident that the FAM scale is too easy for this population. Many subjects ("X" symbols) lie above the most difficult item. Correspondingly, the average difficulty of the items ("M" symbol) is more than 1 SD ("S" symbol) below the average ability of the subjects. This indicates that most of the items were passed by these subjects. We can estimate only roughly their ability. It is like estimating the individual jumping capacities of top athletes, by adopting too low targets, most of which are passed by all of them. Redundancy between many items is also made graphically evident. Out of the 10 residual FAM-specific items 2, only, spread very lightly indeed beyond the range of difficulty encompassed by the FIM^{SM} items. These are Swallowing, which is the easiest item, and Employability, which is the most difficult one.

Table 5, A and B, provides results from a separate analysis of the 13 motor and the 5 cognitive FIM^{SM} items, respectively. A lower number of subjects were misfitting (11 for the motor and 16 for the cognitive subscales, respectively; not shown), although a greater number of subjects got extreme scores (12 and 9, respectively). The range of difficulty covered by both subscales was similar or even wider, compared to the range covered by the 28-item FAM. Again, the average ability of the subjects was more than 1 SD above the average item difficulty, both in the motor and in the cognitive FIM^{SM} subscales, respectively (not shown). In the motor subscale, Tub transfer and Bowel show signs of misfit. None of the cognitive items, by contrast, misfits.

DISCUSSION

The 30-item FAM seems to add nothing to the 18-item FIM^{SM}, as far as TBI outpatients are concerned. Both the FIM^{SM} and the FAM items appear, on average, too easy. This is consistent with the scarce previous literature, showing that the FAM has a marked ceiling effect on TBI outpatients (Hall, Hamilton, Gordon, & Zasler, 1993; Hall, Mann, High, Wright, Kreutzer, & Wood, 1996). Of the 12 new items, two had to be deleted, because of severe misfit, four retained signs of misfit, and eight shared their difficulty levels with the FIM^{SM} items (see also previous reports: Hall, Hamilton, Gordon, & Zasler, 1993; Hall, Mann, High, Wright,

Table 5 (a)
Rasch Analysis of the 13 Motor Items Using the FIM℠ Scale

			INFIT		OUTFIT			
	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	PTBIS	ITEMS
	1.31	.18	1.14	.5	1.74	2.3	.73	T/ShTr
	.87	.19	1.39	1.4	1.24	.9	.71	DressLow
	.70	.19	1.06	.2	.98	-.1	.74	Bath
	.44	.19	.86	-.6	.77	-.9	.83	Toil
	.40	.20	.68	-1.4	.65	-1.5	.86	DressUp
	.40	.20	1.01	.1	.85	-.6	.83	Stairs
	.13	.20	.59	-1.8	.60	-1.7	.89	Walk/Wch
	.09	.20	.90	-.4	.85	-.6	.73	Groom
	-.17	.21	.70	-1.2	.58	-1.7	.88	ToilTr
	-.22	.22	.89	-.4	.73	-1.0	.87	B/C/Wch
	-.73	.24	.95	-.2	.93	-.2	.70	Eat
	-1.22	.26	1.61	1.7	1.71	1.5	.49	Blad
	-2.00	.31	2.40	3.1	1.40	.7	.50	Bowel
Mean	.00	.21	1.09	.1	1.00	-.2		
S.D.	.86	.03	.46	1.3	.38	1.2		

Note. Eleven misfitting and 12 extreme patients were deleted. Person separation 3.25, reliability 0.91; item separation 3.25, reliability 0.91.

Table 5 (b)
Rasch Analysis of the Five Cognitive Items Using the FIM℠ Scale

			INFIT		OUTFIT			
	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	PTBIS	ITEMS
	2.14	.19	1.06	.2	.91	-.3	.60	Mem
	1.05	.21	.63	-1.5	.66	-1.3	.75	PrSolv
	-.07	.25	1.52	1.5	1.21	.7	.71	SocInt
	-.83	.27	.57	-1.8	.57	-1.5	.77	Expr
	-2.28	.33	1.24	.7	.83	-.3	.45	Compr
Mean	.00	.25	1.00	-.2	.84	-.5		
S.D.	1.52	.05	.36	1.3	.22	.8		

Note. Sixteen misfitting and nine extreme patients were deleted. Person separation 2.11, reliability 0.82; item separation 5.37, reliability 0.97.

Kreutzer, & Wood, 1996). Both redundancy and misfit might come out from an intrinsic inconsistency of the 12 new items. In this subscale, both the performance (5 items) and the dependence (7 items) scoring criteria coexist (Tesio, 1997). A relevant number of subjects (17 out of 60) were misfitting. Clinical peculiarities might have strongly biased the scores in such population. Apparently, however, this was not the principal reason for the frequent misfits. Another explanation may be the intrinsic ambiguity of the items. Perhaps the dependence criterion can be unsuitable to measure precisely some high-functioning, nearly independent outpatients. At the same time, in the dependent subjects the relationship between performance and dependence might be quite unpredictable, thus making any given subject or item to appear inconsistent. For instance, a patient might "pass" a difficult item because he reaches independence, while "failing" a much easier item because of a low-quality performance, despite independence. The performance-scored items in the FAM are Emotional status, Adjustment to limitations, Reading, Writing and Attention span. The latter two, only, were fitting. Misfit also affects Community access, a dependence-scored item. In the FAM guidelines (Hall, 1992). Community access "includes the ability to manage transportation, including planning a route, time management, paying fares and anticipating access barriers (excluding car transfer)." This seems quite a multidimensional domain, encompassing many complex motor and cognitive activities, motivation, learning skills etc. In the 28-item FAM scale, one cognitive and one motor FIM^{SM} item, i.e. Social Interaction and Bowel, also misfit. On one hand, psychometric studies have evidenced that for most clinical applications the FIM^{SM} motor and cognitive subscales should be scored independently (Linacre, Heinemann, Wright, Granger, & Hamilton, 1994), and not cumulatively like the FAM requires. On the other hand, the FIM^{SM}, originally conceived for inpatients, also showed intrinsic limitations when applied to this population of TBI outpatients. A Rasch analysis conducted separately on the motor and cognitive FIM^{SM} subdomains (Table 5) still revealed a high number of misfitting subjects. Bowel and Tub retained signs of misfit. The cognitive items, only, became fully fitting.

In conclusion, both the FIM^{SM} and the FAM do not solve the problem of measuring disability in TBI outpatients with a high level of overall independence in basic daily activities. This conclusion needs confirmation from studies on more numerous samples, however. The final calibration of the 28-item FAM was relating to 42 subjects, only. As a premise to future research, one should wonder whether the dependence criterion, even

coherently adopted for all items like in the FIM[SM], should be in itself suitable for this purpose. Rather than burdening and trying to stretch out a disability scale validated in inpatients, perhaps a more rewarding strategy would be one of designing a scale targeted to outpatients and thus mainly scored in terms of performance, not dependence.

REFERENCES

Andiel, C. (1995). Rasch analysis: A description of the model and related issues. *Canadian Journal of Rehabilitation, 9*, 1:17-25.

Granger, C. V., Hamilton, B. B., Linacre, J. M., Heinemann, A.W., & Wright, B. D. (1993). Performance profiles of the Functional Independence Measure. *American Journal of Physical Medicine and Rehabilitation, 72*, 84-89.

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

Hall, K. M. (1992). Functional Assessment Measure. *General rehabilitation and traumatic brain injury program evaluation*. San Jose, CA: Santa Clara Valley Medical Center.

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

Hall, K. M., Mann, N., High, W. M. Jr., Wright, J., Kreutzer, J. S., & Wood, D. (1996). Functional Measures after traumatic brain injury: Ceiling effects of FIM[SM], FIM[SM]+FAM, DRS and CIQ. *Journal of Head Trauma Rehabilitation, 11*, (5):27-39.

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

Linacre J. M., & Wright B. D. (1993). *A User's Guide to BIGSTEPS, Rasch model computer program*. Version 2.4, Chicago: MESA Press.

Ottenbacher K. J., Yungwen H., Granger C. V., & Fiedler R. C. (1996). The reliability of the Functional Independence Measure: A quantitative review. *Archives of Physical Medicine and Rehabilitation, 77*, 1226-1232.

Spinnler H., & Tognoni G. (1987). Italian standardisation and calibration of neuropsychological tests (in Italian). *Italian Journal of Neurological Science*, suppl.8.

Stineman M. G., Escarce, J. J., Goin, J. E., Hamilton, B. B., Granger, C. V., & Williams, S. V. (1994). A case-mix classification system for medical rehabilitation. *Medical Care, 32*, 4:366-379.

Tesio, L., Perucca, L., Franchignoni, F. P., & Porta, G. L. (1996). The effect of age on length of stay, functional independence and discharge destination of rehabilitation inpatients in Italy. *Disability & Rehabilitation, 18,* 10:502-508.

Tesio, L. (1997). Disability, dependence, and performance: which is which? (Editorial). *Europa Medico Physical, 33,* 55-57.

Tsuji, T., Sonoda, S., Domen, K., Saitoh, E., Liu, M., & Chino, N. (1995). ADL structure for stroke patients in Japan based on the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation, 74,* 432-438.

Uniform Data System for Medical Rehabilitation. (1993). UB Foundation, State University of New York at Buffalo,NY. *FIM℠: Functional Independence Measure.* Manuale d'uso. Versione italiana 3.0. Ric. Riabil. (suppl.) 2:pp.1-44, SO.GE.COM Editrice srl, Milano, Italy.

WHO. (1980). *International classification of impairments, disabilities, and handicaps.* Geneva: WHO.

Wright B. D., & Linacre J. M. (1989). Observations are always ordinal; measurement, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70,* 857-860.

Wright B. D., and Linacre J. M. (1994). Reasonable mean-square fit values. *Rasch Measure Trans , 8,* 3:270.

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

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

JOURNAL OF OUTCOME MEASUREMENT®, 2(2), 97- 122

The Effect of Item Pool Restriction on the Precision of Ability Measurement for a Rasch-Based CAT: Comparisons to Traditional Fixed Length Examinations

Perry N. Halkitis

Center for HIV/AIDS Educational Studies and Training
Jersey City State College

This paper describes a method for examining the precision of a computerized adaptive test with a limited item pool. Standard errors of measurement ascertained in the testing of simulees with a CAT using a restricted pool were compared to the results obtained in a live paper-and-pencil achievement testing of 4494 nursing students on four versions of an examination of calculations of drug administration. CAT measures of precision were considered when the simulated examinee pools were uniform and normal. Precision indices were also considered in terms of the number of CAT items required to reach the precision of the traditional tests. Results suggest that regardless of the size of the item pool, CAT provides greater precision in measurement with a smaller number of items administered even when the choice of items is limited but fails to achieve equiprecision along the entire ability continuum.

Requests for reprints should be sent to Perry N. Halkitis, Center for HIV/AIDS Educational Studies and Training, Jersey City State College, 2039 Kenedy Boulevard, Jersey City, NJ 07305

For approximately one century, paper-and-pencil, fixed length examinations have been a mainstay in educational settings. Based on the ideas first documented by E. L. Thorndike (1904), these examinations have provided an opportunity to test large numbers of people efficiently. The popularity of paper-and-pencil examinations throughout this century can be attributed to the relatively weak assumptions of classical test theory, the psychometric model upon which these examinations are based, and the lack of sophisticated technology to handle complex calculations (Hambleton & Jones, 1993). Yet two developments, one technological and the psychometric, have provided the avenue for reconsidering the methods utilized to test individuals (Crocker & Algina, 1986). The availability of large computers has enabled psychometricians to implement item response theory (IRT). This has led to increasingly sophisticated research aimed at improving testing procedures (Kingsbury & Houser, 1993; Weiss & Kingsbury, 1984).

It has been argued that CAT provides a vehicle for measurement that is superior to conventional fixed-length tests simply because it is more efficient and more precise in the determination of the abilities of examinees. The argument of precision is based on the idea that when a test is administered in a computerized adaptive format, individuals are presented items that maximize information at their own ability levels. The result is an individualized or tailored test for each examinee with maximum information, and as a result greater precision in measurement. It is claimed that fewer than half as many questions are needed as in conventional testing, and a CAT yields broad range accuracy in assessing the ability of examinees (Vispoel, Wang & Bleiler, 1997; Ward, 1985). Along with improvements in efficiency, CAT is suggested to provide improved measurement characteristics including measurement precision, reliability, validity, and confidence in pass/fail decisions (Halkitis, 1996; Lord, 1977a; Bergstrom & Lunz, 1993; McBride, 1986; Olsen et al., 1989; Stocking, 1987; Wainer, 1989; Ward, 1985; Weiss, 1985; Weiss & Kingsbury, 1984). In addition to applications in achievement testing (Olsen et al, 1989), computerized adaptive procedures have been implemented for licensure testing (Halkitis & Leahy, 1993; Haynie & Way, 1994), mastery testing (Sheehan & Lewis, 1992), aptitude examinations (McBride, 1986; Schaeffer et al, 1995; Vispoel, Wang & Bleiler, 1997), and diagnostic testing (Tatsuoka & Tatsuoka, 1997). Recent approaches have utilized CAT with open-ended response items (Bennett et al., 1997).

In many cases, Monte Carlo computer simulation studies have been utilized to examine the issue of precision, and the findings of such studies have noted that adaptive tests measure with greater efficiency and with a greater precision when compared to conventional fixed-length examinations (Maurelli & Weiss, 1981; McBride, 1977; Thissen, 1990; Weiss & McBride, 1984). In direct comparisons of CAT with tailored peaked conventional tests, the advantages of CAT were noted (Stocking, 1987). Live testing situations have yielded similar results (Johnson & Weiss, 1980; Lord, 1977;1980; Moreno, et al., 1984; Urry, 1971;1977). In direct comparisons with conventional tests, CAT procedures have been shown to attain the same level of precision using half the number of items (Moreno et al, 1984). More recently, in comparisons of fixed item and computerized adaptive music tests, results indicate the CAT requires 50% to 93% fewer items to match the concurrent validity and reliability of fixed item examinations (Vispoel, Wang, & Bleiler, 1997). Similarly, levels of confidence in pass/fail decisions has been shown to be greater in an a assessment of medical technology among 645 medical technology students when the computerized adaptive test implemented a 90% confidence stopping rule (Bergstrom & Lunz, 1992).

While there has been a substantial amount of research in simulated situations and some live-testing situations, issues regarding CAT require further elucidation. One overriding question concerns the impact of the item pool that is utilized on the effectiveness of a computerized adaptive test. Findings in some investigations have cautioned about the generalizability of results regarding CAT accuracy, noting that indices of accuracy may be limited to "ideal" item pools with rectangular distributions (McBride, 1977), although more recent investigations have noted an enhancement in measurement precision with a pool composed of as few as 278 items (Vispoel, Wang, & Beliler, 1997).

While no specific guidelines exist for the appropriate size and characteristics of item pools, it has been suggested that 100 items might provide satisfactory results for a CAT so long as the item difficulties span the full range of trait levels in the population and items possess high discrimination (Weiss & Kingsbury, 1984), but 150-200 items would be better (Weiss, 1985). Others have noted that the item pool of a CAT requires that the examiner have access to a pool of 200 to 500 items and a database of responses to the items by examinees ranging in number from 300 to 1000, yet without items that span the entire difficulty continuum, increases in measurement quality beyond that of conventional test cannot be assured (De Ayala, Dodd, & Koch, 1990).

This investigation examined the precision of measurement of a computerized adaptive examination with an item pool limited by the realities of the test construction process. For the purposes of the discussion, the item pool is considered limited because it is neither ideal in composition nor theoretically infinite as has been the case in many simulation studies. In this light, this investigation of CAT provides a check regarding the claims of precision being put forth regarding computerized adaptive testing, and speaks to the realities of this testing framework, where unlike in large testing programs, ideal situations are not possible.

METHOD

To assess the advantage in precision of CAT administration in a situation when an item pool is limited in nature, measures of precision defined as standard error of ability measurement [SEM] in the Rasch model were used as a basis of comparison (Wright & Douglas, 1975):

$$(\theta)^2 = \frac{d^2}{d\theta^2}[\ln L(u\,|\theta)].$$

In the first comparison, standard errors of measurement ascertained in the paper-and-pencil administrations were compared to those ascertained in a simulated CAT situation where test length was held constant using a Monte Carlo simulation. In the second assessment, the number of items required for the CAT with this limited item pool to match the SEM's obtained through the fixed-length paper-and-pencil administrations was determined. The goal of these analyses was to explore the differences in precision achieved via these two testing frameworks given the limited nature of the item pool that was utilized.

Ability estimates and measures of precision were ascertained for a randomly selected sample of 4494 examinees who were administered one of four versions of *Calculations of Drug Administration*, a pilot achievement examination designed for registered nursing students who had completed the appropriate course preparation in pharmacology principles at accredited schools at the time of test administration in February, 1992. The examination was part of a larger exam of pharmacology which consisted of the Calculations section and two additional sections, *Principles of Medication Administration* and *Side Effects of Medications* (National

League for Nursing, 1993). Initial analyses of the entire examination revealed three factors; the Calculations subtest represented one of these factors and is the examination that was utilized in this investigation.

Examinees were students in randomly selected schools of nursing in the United States. Four versions of the Calculations examination each consisting of 30 unique items and linked by a set of fifteen anchor items were utilized to allow for the experimentation of a larger set of items.

Examinee responses were calibrated for the 1-PL model using BILOG 3 (Mislevy & Bock,1990). Initial calibrations of item difficulty and fit statistics were examined to assess the overall fit of the items to the model. Items judged as misfitting were then eliminated from the item pool (Halkitis, 1992; Hambleton, Swaminathan & Rogers, 1991).

Paper-and-Pencil Administrations

The final pool of 101 items were recalibrated using BILOG. Indices of fit of items were assessed to assure that the items actually fit the 1-PL Rasch model (Hambleton, 1989; Wright & Douglas, 1975). The unidimensionlity of the Calculations examination also was assessed resulting in one primary factor of eigenvalue 4.2 (Halkitis, 1993). Mean-square statistics were computed for the items as indices of fit and conjunction with the power of the hypothesis test were judged as either fitting or misfitting (Halkitis, 1992). In addition, estimates of ability and standard errors of measurement were calibrated for each examinee pool. The final item pool utilized in these calibrations consisted of twelve anchor items and 89 unique items. The four versions of the examination consisted of 33, 33, 35, and 36 items respectively, linked by a set of twelve common items. The ability estimates, standard errors of measurement, test information function, and test length provided the basis of comparison with the CAT simulated data.

Simulated CAT Administration

The CAT item pool consisted of the 101 unique items calibrated on the fixed-length paper-and-pencil examinations. Item difficulty calibrations based on paper-and-pencil administrations have been shown to be equivalent to CAT calibrations (Bergstrom & Lunz, 1991; Hetter, Segall & Bloxom, 1994). The entire pool was however limited in its composition, due to both number of items and item difficulty distributions. The pool was neither infinite nor ideal in its composition: mean item difficulty for

the pool was -0.143 logits (s=0.566) with a maximum and minimum item difficulty of 1.711 and -1.362 respectively (Halkitis, 1995).

Simulees were drawn from a hypothetical uniform, rectangular distribution (-3.00, +3.00) at twelve critical points (-1.00 to +1.75 inclusive, at intervals of 0.25 logits along the continuum). The critical range was chosen as it represented the ability continuum ascertained in the paper-and-pencil administrations. Fifty simulees were generated at each of the critical points along the continuum to simulate an examinee pool of 600 students. For each, a maximum of 36 responses were generated so to match the maximum test length of the longest paper-and-pencil administration. Ability estimates and standard errors of ability measurement were noted after the adaptive administration of each item. This data provided the basis for comparisons of precision at the critical points in each of the four traditional test administrations.

Responses were simulated using a random number generator (Halkitis, 1995) written in the programming language *C++* (Borland, 1993). After the determination of the next item to be administered based on maximizing information (i.e. the item with difficulty that most closely matched ability estimate), $P(x=1 \mid \theta)$ was calculated. To determine if the response was correct or incorrect, a random number was generated; if this number fell within the probability of $P(x=1/\theta)$, then the response was determined to be correct; else, it was incorrect. This response was entered into a *Paradox v. 4.0* (Borland, 1990) database where the new estimated ability and standard error were calculated using a ML approach (Halkitis, 1995). The same procedure was undertaken for the administration of each of the 36 items where the presumed latent ability of the examinee in conjunction with the item difficulty was used to determine the probability of correct response, and ML procedures were used to estimate ability of the simulees.

To initiate the estimation of ability and standard error and to allow for the convergence of the ML estimates, two responses (one correct, one incorrect for two items) were presupposed for each simulee. The items were each assumed to have a difficulty parameter of 0.00 resulting in an initial ability estimate of for each simulee. Given no prior information, 0.0 presented a "best-guess" estimate of ability for the presupposed uniform distribution of -3.00 to +3.00. The first "true" item presented was of difficulty 0.01 as this difficulty was closest to the presumed average ability of the simulee pool. Three such items were part of the 101 item pool. Establishing an initial estimate of ability based on ability means has been used elsewhere (Vispoel, Wang & Bleiler, 1997).

The step by step adaptive procedure utilized was as follows:

1. Set the initial estimate of θ equal to 0.0 by assuming the presentation of two false items of difficulty 0.0, one marked as correct response, the other as incorrect response. These two false items were dropped from the estimation after the administration of fifteen "true" items.

2. Present an item from the pool closest to 0.0 logits; in this case an item of difficulty 0.1 logits.

3. Flag the item as presented, so that it will not be administered again to the same examinee.

4. Determine the probability of a correct response given examinee estimated ability and item difficulty.

5. Randomly generate a number from .01 to 1.00. Random number generation was undertaken in an ancillary program written in *C++* (Halkitis, 1995).

6. If randomly generated number is less than or equal to the probability of correct response, mark the response as correct; else mark the response as incorrect.

7. Enter response, as either correct or incorrect, into Paradox database and recompute ability estimate.

8. Select as the next item for presentation the item whose difficulty is the smallest absolute distance from newly estimated ability. Absolute distance is measured as the difference in logit units between the item difficulty and estimated ability.

9. Repeat steps 3 through 7 until 36 items are presented.

Comparing Testing Frameworks

In assessing the advantages of a CAT methodology, the data were examined using two approaches. In the first analysis, the precisions of ability measurement (SEMs) ascertained in the CAT administration of a fixed number of items were compared to the standard errors of measurement obtained in each of the fixed length paper-and-pencil administration at each of the thirteen ability points along the continuum assuming rectangular distribution of ability estimates. Comparisons of the standard errors of measurement at these fixed points provided a method for assessing the precision of the two approaches after a given numbers of items were administered and allowed a judgment to be made regarding the precision that is achieved.

In the second analysis, the number of items required by the CAT procedure to achieve the level of precision (SEM) ascertained in the paper-and-pencil administration was computed. Comparisons of the number of items required to achieve this equiprecision provided a basis for assessing the advantages of the CAT administration.

For the purposes of the comparisons, examinees answering all questions correctly or incorrectly on the paper-and-pencil administrations were not utilized as ML estimation would provide less than ideal measures for these examinees. Further, estimations for the CAT simulation incorporated an assumption of at least one incorrect response at mean ability level to initiate calibration. The final comparison group was composed of 1185, 992, 1097, and 1097 (N = 4371) examinees on Forms A to D respectively. Ability estimates range from a low of -0.98 on Form B to a maximum ability estimate of 1.79 on Form C. For that reason, the critical points of comparison were chosen to range from -1.00 to +1.75 from a uniform distribution of simulees.

RESULTS

SEM Comparisons

A direct comparison of CAT and paper-and-pencil results was undertaken by comparing the SEMs achieved after the administration of *n* adaptive items to *n* length paper-and-pencil tests. Thus, the SEM at each critical point along the ability continuum was compared to the SEM ascertained at each of these ability levels at the end of the paper-and-pencil tests. Tables 1 through 3 provide a summary of the mean SEM at each of the critical points and the mean ability estimate for a 33, 35, and 36 item adaptive tests respectively, as well as the SEM for the paper-and-pencil administrations. A visual inspection of the data indicates that for every ability level as well as every test length the SEM achieved in the CAT simulation is smaller than that achieved in the paper-and-pencil administration. On the 36 item (Form D) comparison, SEM ranged from .20 (b=-0.25) to .35 (b=1.75) on the paper-and-pencil examination, while for the corresponding abilities on the CAT, SEM ranged from a low of .192 to a high of .32. For each of the twelve critical points, the SEM ascertained via the CAT administration was lower than that achieved through paper-and-pencil administrations of the same item length. The smallest difference existed about the center of the ability continuum where the majority of the items are

Table 1
Percision of 36-Item Adaptive Test and 36-Item Conventional Test (Form D)

Latent Ability	Paper-and-Pencil Standard Error Form D	Simulated CAT Standard Error (mean)	Simulated CAT Ability Estimate (mean)
-1.00	.23	.207	-1.031
-0.75	.22	.201	-.774
-0.50	.21	.197	-.499
-0.25	.20	.192	-.246
+0.00	.21	.198	-.130
+0.25	.21	.197	.254
+0.50	.22	.199	.506
+0.75	.23	.213	.757
+1.00	.26	.221	.989
+1.25	.28	.244	1.221
+1.50	.31	.274	1.525
+1.75	.35	.320	1.763

Table 2
Percision of 35-Item Adaptive Test and 35-Item Conventional Test (Form C)

Latent Ability	Paper-and-Pencil Standard Error Form C	Simulated CAT Standard Error (mean)	Simulated CAT Ability Estimate (mean)
-1.00	.23	.210	-1.015
-0.75	.22	.202	-.799
-0.50	.21	.200	-.532
-0.25	.21	.200	-.228
+0.00	.21	.200	-.040
+0.25	.21	.200	.246
+0.50	.23	.203	.533
+0.75	.24	.214	.769
+1.00	.26	.221	.978
+1.25	.29	.245	1.219
+1.50	.32	.280	1.528
+1.75	.34	.320	1.755

Table 3
Percision of 33-Item Adaptive Test
and 33-Item Conventional Test (Forms A & B)

Latent Ability	Paper-and-Pencil Standard Error Form		Simulated CAT Standard Error (mean)	Simulated CAT Ability Estimate (mean)
	A	B		
-1.00	.24	.23	.217	-1.021
-0.75	.23	.22	.208	-.786
-0.50	.21	.21	.206	-.504
-0.25	.21	.21	.200	-.242
+0.00	.20	.21	.204	.000
+0.25	.21	.22	.201	.246
+0.50	.22	.23	.210	.593
+0.75	.24	.25	.222	.805
+1.00	.27	.27	.228	.956
+1.25	.29	.30	.248	1.204
+1.50	.33	.34	.284	1.510
+1.75	.36	.37	.321	1.736

clustered. Similar results are indicated for the 35 and 33 item comparisons demonstrated in Tables 2 and 3 respectively.

The gains in accuracy noted above are realized in Figures 1 through 4 which depict the SEM functions of the *n* length adaptive test along with their *n* length paper-and-pencil counterparts. (Note that two such graphs are provided for the 33 item CAT as there are two 33 item paper-and-pencil tests to which to compare.) In each instance, the function of the CAT depicts greater accuracy along the entire ability continuum. In addition, the functions of the CATs were all flatter than those of the paper-and-pencil administrations, indicating that not only was the CAT more informative and accurate along the entire ability continuum but also that precision estimates about the entire continuum were more disparate in the pa-

per-and-pencil administrations than the CAT. Information was maximized about the mean item difficulty for the paper-and-pencil tests and was less pronounced in ability regions where the number of items matching the ability is limited. This would confirm earlier notions that a CAT is more accurate in its estimation of ability and that this accuracy tends to be more equivalent between abilities on a CAT administration than on a conventional exam.

A numerical comparison of SEMs achieved in the conventional testing to those ascertained in CAT simulation is given in Table 4. The difference between the mean CAT SEM and mean paper-and-pencil SEM for the critical ability levels is provided in the first four columns for the 35 item, 36-item, and two 33 item exams. This is followed by four columns indicating a proportional comparison of CAT SEM over paper-and-pencil SEM. For the 33 item examinations, the greatest difference in precision between the two testing frameworks was noted at the upper end of the ability continuum (1.25 to 1.75), where SEM was approximately 20% smaller on the CAT than on the paper-and-pencil exam. A similar pattern was noted for the 35 and 36 item exams. This phenomenon can be explained by considering the item difficulty distributions of the paper-and-pencil exams. On all four paper-and-pencil exams, items of higher difficulty level were under represented.

Non-Uniform Distributions. Discussion for the results thus far has been based on the assumption of a uniform distribution of ability. Results also were considered in light of non-uniform distributions to determine the advantage of CAT with limited item pools in those more realistic ability distributions. To undertake these analyses, five non-uniform, normal distributions were considered (0,1), (.5,1), (1,1), (-.5,1), (-1,1).

Mean SEM's associated with each of the critical ability points in the non-uniform distribution were determined by weighting mean SEM's associated with each critical ability level in the rectangular distribution by the probabilistic area associated with each critical ability point in each of the non rectangular distributions. Probabilistic areas associated with each of the critical points in the non-uniform distribution were determined based on ability bands about the critical point taken as the distance half-way between the critical point and the adjacent upper and lower critical points. Thus, the mean SEM associated with the ability in the uniform distribution was weighted according to the area of the non-uniform distribution associated with each of the ability bands. In the case of the uniform distribution N(0,1), the SEM for ability level 0.0 was weighted 0.103 corre-

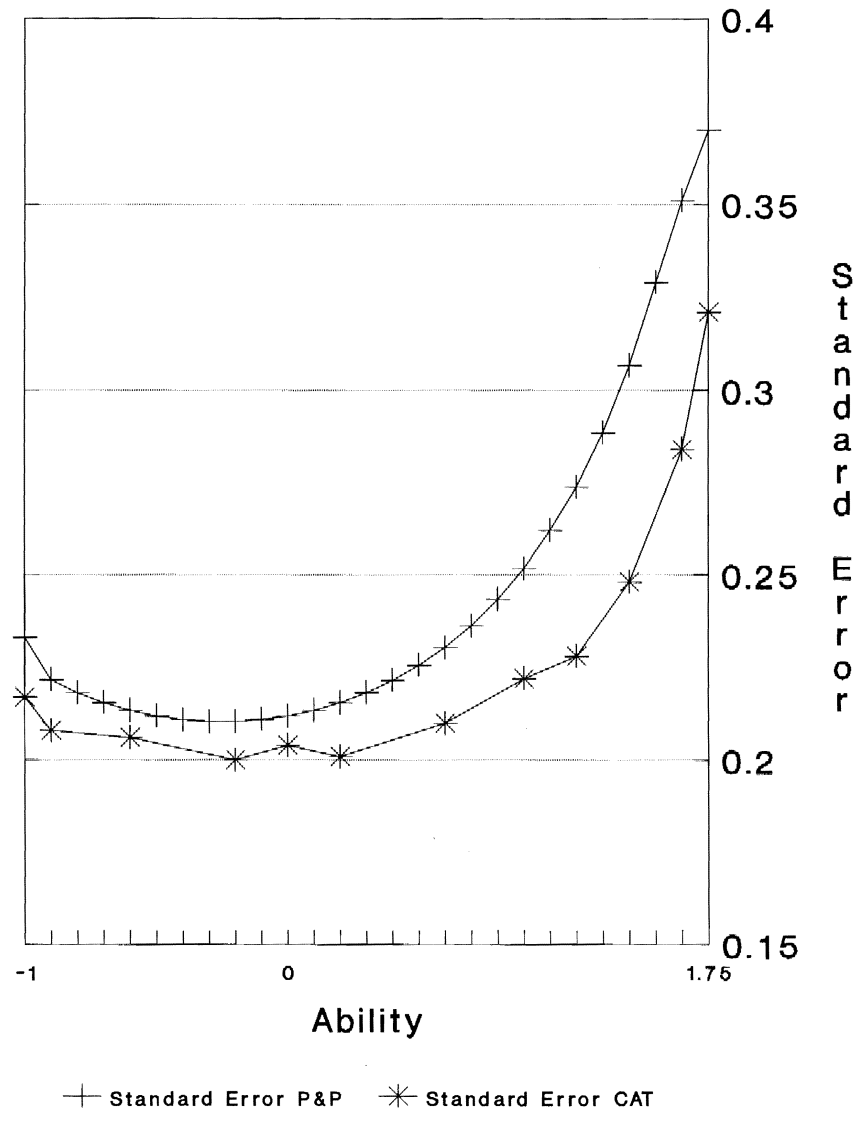


FIGURE 1 Standard errors CAT vs. paper-and-pencil A (33 items).

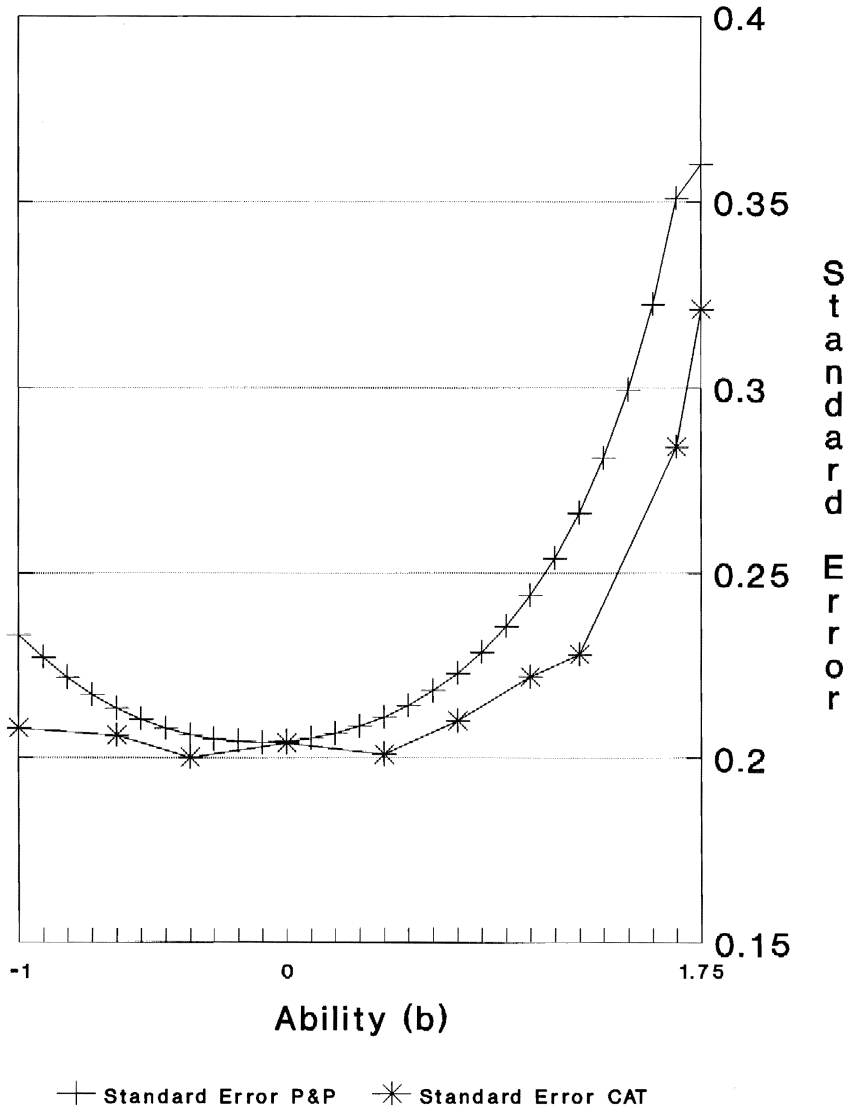


FIGURE 2 Standard errors CAT vs. paper-and-pencil B (33 items).

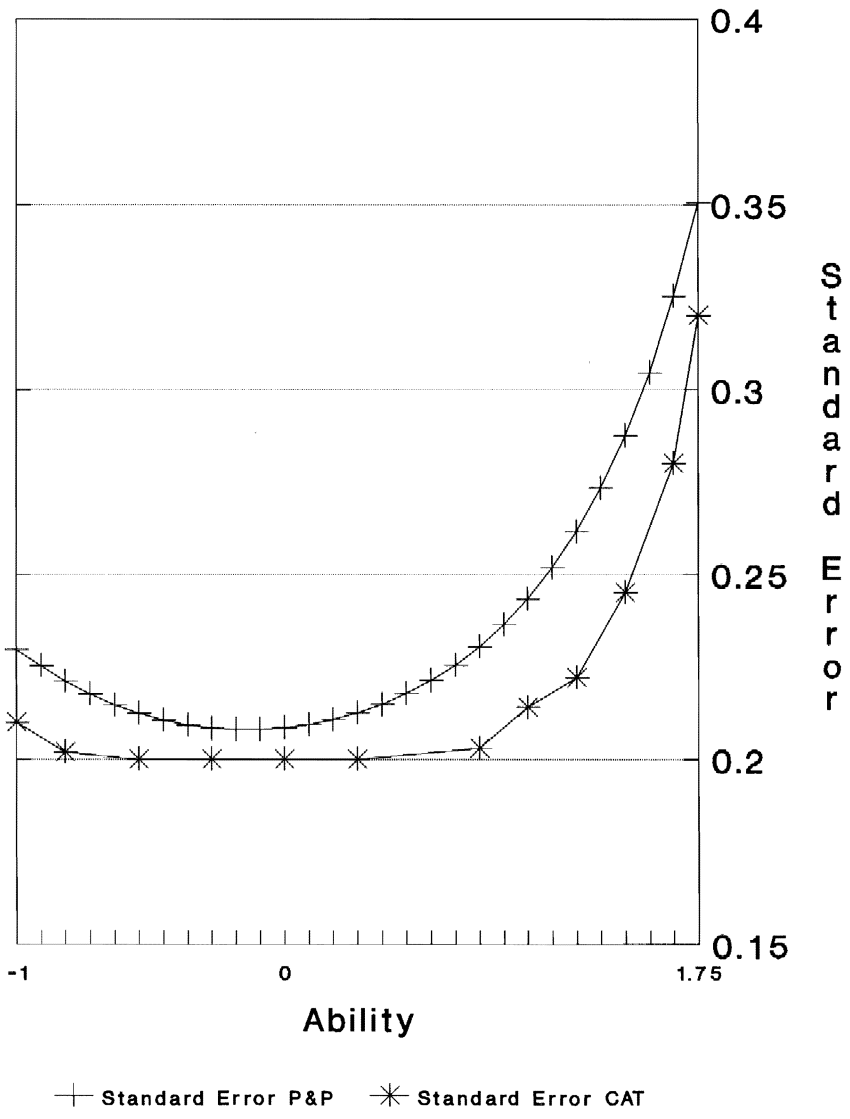


FIGURE 3 Standard errors CAT vs. paper-and-pencil C (35 items).

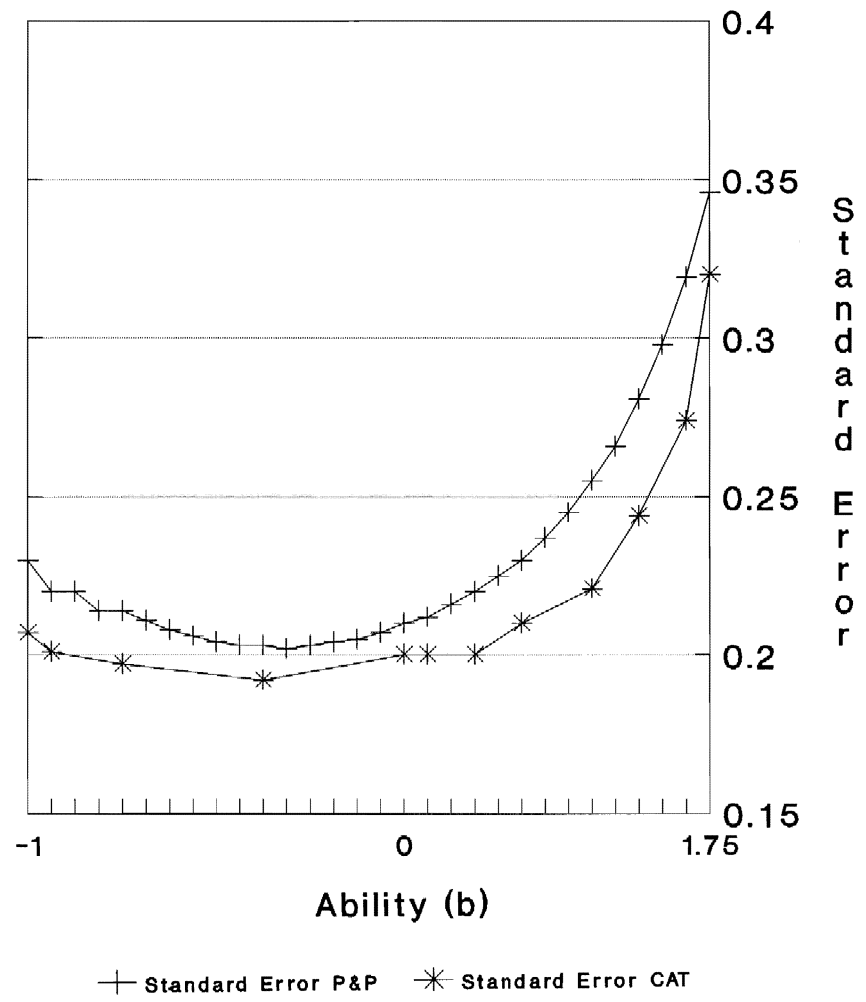


FIGURE 4 Standard errors CAT vs. paper-and-pencil D (36 items).

Table 4
Comparison of Paper-and-Pencil
and CAT Standard Errors of Measurement for a Uniform Ability Distribution

Latent Ability	SEM DIFFERENCE (SEM_{CAT}-$SEM_{P\&P}$)				$(SEM_{CAT})^2/(SEM_{P\&P})^2$			
	33 Items		35 Items	36 items	33 Items		35 Items	36 Items
	A	B	C	D	A	B	C	D
-1.00	-.023	-.013	-.020	-.023	.817	.890	.834	.810
-0.75	-.022	-.012	-.018	-.019	.817	.894	.843	.835
-0.50	-.004	-.004	-.010	-.013	.962	.962	.907	.880
-0.25	-.010	-.010	-.010	-.008	.907	.907	.907	.922
0.00	-.004	-.006	-.010	-.012	.980	.944	.907	.889
+0.25	-.009	-.019	-.010	-.013	.916	.835	.907	.880
+0.50	-.010	-.020	-.027	-.012	.911	.834	.779	.818
+0.75	-.018	-.028	-.026	-.017	.856	.789	.795	.858
+1.00	-.042	-.042	-.038	-.039	.713	.713	.729	.723
+1.25	-.042	-.052	-.045	-.036	.731	.713	.714	.759
+1.50	-.046	-.056	-.040	-.036	.741	.698	.765	.781
+1.75	-.039	-.049	-.042	-.030	.795	.753	.886	.836

sponding to the area associated with this ability band (-0.125 to 0.125) in the distribution; the mean SEM for ability level 0.25 was weighted .096 representing the for the ability band about 0.25 (i.e., -.125 to 0.375). Ability bands were calculated for each of the other critical ability points in the same manner using the calculation: Mean SEM (associated with critical ability level in rectangular distribution) * probabilistic area (associated with the critical ability band in the distribution). The weighted sum thus provided the mean SEM for the distribution to be used in the comparison with the SEMs derived from the fixed-length traditional examinations. For the other Normal distributions, N(.5,1), N(1,1), N(-.5,1), N (-1,1), the same procedure was utilized adjusting for the mean value in each of the distributions.

Mean SEMs were thus computed for each of the non-uniform distri-

butions and compared to the mean SEMs achieved in the fixed length tradition administrations. Table 5 provides the mean gains in accuracy when considering the uniform and four non-uniform distributions. When CAT measures of precision were compared to these indices on the paper-and-pencil exams, the greatest gains were noted when we assume the distribution of ability to be Normal (1,1), i.e. high ability examinees. A mean gain of .040 SEM was achieved when the 33-item CAT is compared to the 33-item paper-and-pencil exam (A); the gain was .040 SEMs when the CAT was compared to version B (33 items), .045 SEM when compared to the 35-item paper-and-pencil exam, and 0.37 SEM when compared to the 36-item paper-and pencil exam. Not surprisingly, the N(-.5, 1) and N(-1,1) distributions demonstrated the least amount of accuracy enhancement when the paper-and-pencil exams were compared to the CAT. For the N(-5,1) distribution mean gains ranged from .015 to .018 and for the (-1,1) distributions gains ranged from .022 to .026. These gains in accuracy of the CAT over the paper-and-pencil exams are even less than those attained when we assumed a uniform distribution.

Test Length Comparisons

Table 5
Mean Gain in Precision (SEM) Using CAT for Theoretical Uniform and Non-Uniform Ability Distributions

	Theoretical Distributions					
	Uniform	Normal (-1,1)	Normal (-.5,1)	Normal (0,1)	Normal (.5,1)	Normal (1,1)
33-item CAT (compared to A)	.022	.016	.022	.031	.036	.040
33-item CAT (compared to B)	.026	.015	.023	.034	.044	.050
35-item CAT	.025	.018	.026	.032	.042	.045
36-item CAT	.022	.018	.024	.030	.035	.037

In the second analysis of the results, the data were examined by comparing the number of adaptive items need to achieve the same precision as each *n*-length paper-and-pencil tests. Thus, the data were analyzed by determining the mean SEM after the administration of *n* traditional items and then determining at which point along the adaptive simulation each examinee ascertained this SEM. The mean number of CAT items required to achieve the same precision as each of the paper-and-pencil exams at each of the critical points are listed in Table 6.

When compared to paper-and-pencil exam version A (33 items), the CAT required a smaller number of items to achieve the same level of precision as the traditional administration. Only at a latent ability of 0.00 was the mean number of items the same. Overall, the CAT required a mean number of 25.6 items as compared to the 33 items of the paper-and-pencil exam. Again, the results were most dramatic at the extreme ability levels, where the number of available items was most limited. At an ability of 1.75 the 33 item paper-and-pencil exam (Form A) achieved a precision of .36 SEM; this level of precision was achieved with a CAT of 18.6 items. A similar result, although less extreme, was evidenced at ability level -1.00, where only 25.2 items was required to achieve the precision of the 33 item paper-and-pencil exam. In the middle of the ability/difficulty continuum where there were larger number of items on the paper-and-pencil exams, the number of CAT items required to achieve equiprecision was also less, but not dramatically less. Similar results were achieved when the CAT administration was compared to the other 33- item paper-and-pencil administration (Form B), 36-item paper-and-pencil exam (Form D), and the 35-item traditional administration (Form C). When one considers the overall mean number of CAT items needed to achieve equiprecision about the entire ability continuum, eight to ten less items administered were required to achieve the same level of precision as the paper-and-pencil administrations. Approximately 25 to 26 items were required by the CAT to achieve the same level of precision of the 33 item traditional examinations; approximately 26 items were need to achieve the same precision level as the 35 item paper-and-pencil exam; and approximately 27 items were required by the CAT to achieve the same level of precision as the 36-item paper-and-pencil exam. This confirmed earlier notions that CAT achieves equal levels of precision as traditional paper-and-pencil tests with a smaller number of items administered.

Table 6
Mean Number of Items Required to Achieve Equipersision withCAT as Compared to Paper-and-Pencil Administrations

Latent Ability	SEM Form A(33)	Items Needed by CAT	SEM Form B(33)	Items Needed by CAT	SEM Form C(35)	Items Needed by CAT	SEM Form D(36)	Items Needed by CAT
-1.00	.24	25.2	.23	27.4	.23	27.4	.23	27.4
-0.75	.23	26.6	.22	28.8	.22	28.8	.22	28.8
-0.50	.21	30.8	.21	30.8	.21	30.8	.21	30.8
-0.25	.21	29.3	.21	29.3	.21	29.3	.20	32.4
0.00	.20	33.2	.21	30.2	.21	30.2	.20	33.2
+0.25	.21	29.8	.22	27.2	.21	29.8	.21	29.8
+0.50	.22	28.6	.23	25.6	.23	25.6	.22	28.6
+0.75	.24	26.2	.25	24.7	.24	26.2	.23	28.3
+1.00	.27	19.6	.27	19.6	.26	21.3	.26	21.3
+1.25	.29	18.9	.30	17.1	.29	18.8	.28	19.2
+1.50	.33	19.6	.34	17.6	.32	20.0	.31	22.0
+1.75	.36	18.6	.37	15.7	.34	20.1	.35	19.1
MEAN		25.5		25.7		25.7		26.7

DISCUSSION

In order to determine the effectiveness of adaptive testing in situations where item pools are limited, a simulation study was conducted comparing the effectiveness of the CAT with a pool of 101 items to four paper-and-pencil exams consisting of 33, 35, or 36 items. The ability estimates of simulees, with presumed latent abilities ranging from -1.00 to +1.75, were generated along with the standard errors of measurement after the administration of 33, 35, and 36 tailored items. The items were drawn from the 101 without replacement. Accuracy estimates were compared to those attained in a live testing situation of 4494 examinees who took one of the four versions of the paper-and-pencil examination. Comparisons of the two testing frameworks were undertaken by considering the SEMs for the paper-and-pencil exams as compared to those generated in the adaptive framework. In addition, the number of adaptive items required to achieve the accuracy level of the paper-and-pencil exams was determined. Initial comparisons were undertaken assuming an a uniform ability distri-

bution ranging from -1.00 to +1.75 logits.

In both of the above sets of analyses, the CAT proved to be superior to the paper-and-pencil administration both in terms of accuracy at the conclusion of the administration and in terms of the quantity of items required for the CAT to reach the level of accuracy of the traditional administrations. This enhanced accuracy of the adaptive examinations at each of the ability levels over the paper-and-pencil examinations was confirmed statistically through the non-parametric sign test and was also evidenced when the distributions of ability were assumed to be non-uniform in nature.

While the overall comparisons of the adaptive testing to the non-adaptive testing suggest greater precision in measurement by the adaptive exam, the results also depict an inequality of precisions about the entire ability continuum. Like the paper-and-pencil examinations, which lacked items at the high end of the ability continuum resulting in higher SEMs at these abilities, so too the adaptive examinations failed to achieve equiprecision about the entire continuum with much lower measures of accuracy being ascertained at levels greater than +1.25. While the SEM function curve is lower than that of its paper-and-pencil counterparts at these abilities, the adaptive exam still fails to achieve a precision comparable to those achieved at the lower ability levels. Thus it might be suggested that while the adaptive examination, even with its limited pool, does enhance accuracy when compared to non-adaptive administration, it is not effective in achieving equiprecision about the entire ability continuum when the item pool is limited in its composition.

While the above results would clearly establish the superiority of adaptive testing to paper-and-pencil testing, where all examinees takes the same set of items, it does nonetheless demonstrate the limitations of the approach. Like all tests that are as good as the items that constitute them, the effectiveness of CAT is dependent upon the item pool from which questions can be drawn. And while the adaptive test by its very nature of tailoring is sure to yield smaller SEMs, the framework in and of itself cannot provide an ultimate solution for accuracy of measurement at ability levels where these is a lack of items of equivalent difficulty. This result is clearly demonstrated by this study at the upper end of the ability continuum where the CAT provides greater accuracy than the paper-and-pencil exam, but still fails to achieve equiprecision to lower ability levels. It might, in fact, be argued that the sole reason that the CAT does provide a greater level of accuracy at these extreme abilities because items from all four paper-and-pencil forms are pooled into a larger set of items from which the CAT is

administered. In other words, the most difficult items from all four paper-and-pencil exams are available for administrations. Conversely, on average only one-quarter of these difficult items are available on each of the four paper-and-pencil exams, thus the increased accuracy of the CAT at these levels is not surprising. In fact, this is the very reason that increased accuracy is achieved at all of the examined ability levels, and the reason that the comparison that assumes an ability distribution (1,1) demonstrates the greatest advantage of the adaptive over the non-adaptive testing.

One might argue that this is one of the basic tenets of adaptive testing--to pool all available items and select from among them as needed for each examinee, as compared to selecting n items to create an n-length paper-and-pencil test. The true test of the adaptive framework in its ability to overcome the limitations of a restricted pool, however, is not noted. The CAT parallels the results of the traditional tests in the shape of the information function, but simply increases its height due to the fact that the items are tailored and a larger selection of items is available for administration. These results corroborate previous findings (Stocking, 1987) where 20-item adaptive exams were compared to 20-item non-adaptive exams utilizing a variety of item difficulty distributions.

Nonetheless, the adaptive testing framework does enhance accuracy along the entire ability continuum, albeit the limited nature of the item pool prevents the theoretical equiprecision about the entire continuum to be achieved. Even when there are only 101 items from among which to choose, adapting an examination via mechanisms such as CAT will allow for greater precision with a smaller number of items than a traditional exam where the same subset of items from the pool are administered to all examinees. The limited nature of the item pool simply limits the ability to achieve equiprecision, not the ability of CAT to achieve superiority of traditional non-adaptive tests. Enhanced accuracy with a smaller number of items is still noted in the adaptive framework.

The results suggested above are based on the live testing of examinees on a paper-and-pencil examinations and on simulated candidates in the adaptive testing framework. Perhaps, further evidence for the arguments set above would be achieved if actual examinees were tested in both frameworks, thus allowing for a comparison of within the adaptive and non-adaptive frameworks. Simulated examinee abilities are based solely upon probability theory and do not allow room for human factors which, in fact, could be present in CAT administrations. Further investigations along this line might implement live candidates for both frameworks.

Finally, each of the four paper-and-pencil exams were randomly created so to adhere to content fit. In the pilot testing of the examinations, no consideration was given to creating exams with particular difficulty distributions as was done in previous studies (Stocking, 1987). Thus further comparisons might be made between paper-an-pencil exams and CATs with limited pool, when the paper-and-pencil exams are designed to adhere to specific item difficulty distributions.

In the end, work in the area of computerized adaptive testing needs to continue. Clearly, the negligible effect of the medium of presentation has been established (Mead & Drasgow, 1993) as has the superiority of CAT in terms of accuracy and efficiency in hypothetical situations, as well as the ability of CAT to achieve these desirable psychometric qualities even when item pools are limited. As CAT becomes more widely implemented, practical shortcomings and issues need to be clearly documented and investigated so to assure that this shift in testing frameworks ultimately provides solutions to the traditional testing. Recent controversies such as the memorizing of items by candidates administered the Graduate Record Examination via CAT necessitate the consideration of such issues if there will ever be a large enough item pool to counteract cheating.

Still despite its limitations, CAT holds the potential for a testing process that more accurately assesses the width and breath of examinees' abilities (Cooper & Halkitis, 1995).

ACKNOWLEDGMENTS

The author wishes to acknowledge Drs. David Rindskopf, Alan Gross, and Roger Millsap for their support of this project and Dr. Richard Smith for his inspiration.

REFERENCES

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an automatically scorable open-ended response type for measuring mathematical reasoning in computer-adaptive tests. *Journal of Educational Measurement, 34*(2), 162-176.

Bergstrom, B. A., & Lunz, M. E. (1992). Confidence in pass/fail decisions for computer adaptive and paper and pencil examinations. *Evaluation & the Health Professions, 15*(4), 453-464.

Bergsstrom, B. A., & Lunz, M. A. (1991, April). *Equivalence of Rasch item calibrations and ability estimates across modes of administration.* Paper presented

at the annual International Objective Measurement Workshop, Chicago, IL.

Borland International (1993). *C ++ version 4.0.* Scotts Valley, CA: Borland International.

Borland International (1985). *Paradox version 4.0.* Scotts Valley, CA: Borland International. Borland International (1993). *C ++ version 4.0.* Scotts Valley, CA: Borland International.

Cooper, C., & Halkitis, P. N. (1995). This test is for you. *Wired, 3*(1), 64-68.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Holt, Rinehart, and Winston Inc.

De Ayala, R. J., Dodd, B. G., & Koch W. R. (1990). A simulation and comparison of flexilevel and Bayesian computerized adaptive testing. *Journal of Educational Measurement, 27*(3), 227-240.

Halkitis, P. N. (1996). CAT with a limited item bank. *Rasch Measurement, 9*(4), 471.

Halkitis, P. N. (1995). *An examination of the precision of measurement of computerized adaptive test with limited item pools.* Unpublished doctoral dissertation, Graduate Center of The City University of New York.

Halkitis, P. N. (1993). Personal correspondence.

Halkitis, P. N. (1992). Mean square significance and sample size. *Rasch Measurement, 6*(3) 227-228.

Halkitis, P. N., & Leahy, J. M. (1993). Computerized adaptive testing. *Nursing & Health Care, 14*(7), 378-385.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory: Volume 2.* Newbury Park, CA: Sage Publications.

Haynie, K. A., & Way, W. D. (1994, April). *The effects of item pool depth on the accuracy of pass/fail decisions for the NCLEX using CAT.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

Hetter, R. D., Segal, D. O., & Bloxom, B. M. (1994). A comparison of item calibration media in computerized adaptive testing. *Applied Psychological Measurement, 18*(3), 197-204.

Johnson, M. F., & Weiss, D. J. (1980). Parallel forms reliability and measurement accuracy comparison of adaptive and conventional testing strategies. In D. J. Weiss (Ed.) *Proceedings of the 1979 computerized adaptive testing conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Kingsbury, G. G., & Houser, R. L (1993). Assessing the utility of item response models: Computerized adaptive testing. *Educational Measurement: Issues and Practice, 12*(1) 21-27.

Lord, F. M. (1977a). A broad range tailored test of verbal ability. *Applied Psycho-*

logical Measurement, 1, 95-100.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*(2), 117-138.

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

Maurelli, V., & Weiss, D. J. (1981). *Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries* (Research Rep. No. 81-4). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

McBride, J. R. (1986). *A computerized adaptive editions of the Differential Aptitude Tests.* New York: Psychological Corporation. (ERIC Document Reproduction Service No. ED 285 918).

McBride, J. R. (1977). Some properties of Bayesian adaptive ability testing strategy. *Applied Psychological Measurement, 1*(1), 121-140.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114, 3*, 449-458.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3, Item analysis and test scoring with binary logistic models.* Chicago: Scientific Software.

Moreno, K. E. et al. (1984). Relationship between corresponding armed services vocational aptitude battery (ASVAB) and Computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement, 8*(2), 155-163.

National League for Nursing (1993). *CAT pharmacology test.* New York: Author.

Olsen, J. N., Maynes, D. D., Slawson, D., & Ho, K. (1989). Comparison of equating of paper-administered, computer-administered, and computerized adaptive achievement tests. *Journal of Educational Computing Research, 5*(3) 311-326.

Schaefer, G. A., Steffen, M., Golub-Smith, M. L., Mills, C. N., & Durso, R. (1995). *The introduction and comparability of the computer-adaptive GRE General Test* (Research Report No. RR-95-20). Princeton NJ: Educational Testing Service.

Sheehan, K., & Lewis. C. (1992). Computerized adaptive tetsing with non-equivalent testlets. *Applied Psychological Measurement, 16*, 65-76.

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review, 36*(3/4), 263-277.

Tatsuoka, K. K., & Tatsuoka, M.M. (1997). Computerized cognitive diagnostic adaptive testing: Effects on remedial instruction as empirical validation. *Journal of Educational Measurement, 34*(1), 3-20,

Thissen, D., & Mislevy, R. J. (1990). Item response theory, item calibration and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Earlbaum Associate, Inc.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements.* New York: Science Press.

Urry, V. W. (1977). Tailored testing: A successful application of item response theory. *Journal of Educational Measurement, 14*(2), 181-196.

Urry, V. W. (1971). *Individualized testing by Bayesian estimation* (Research Bulletin 0171-177). Seattle: University of Washington, Bureau of Testing.

Vispoel, W. P., Wang, T., & Bleiler, T. (1997). Computerized adaptive and fixed-item tetsing of music listening skill: A comparison of efficiency, precision, and concurrent validity. *Journal of Educational Measurement, 34*(1), 43-63.

Wainer, H. (1989). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ward, W. C. (1985). Measurement research that will change test design for the future. In *The Redesign of Testing for the 21st Century, Proceedings of the 1985 Invitational Conference*. Princeton, NJ: ETS.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53, 6,* 774-789.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21, 4,* 361-375.

Weiss, D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8, 3,* 273-285.

Wright, B. D., & Douglas, G. A. (1975). *Best test design and self-tailored testing*. Research Memorandum No. 19, Statistical Laboratory, Department of Education, University of Chicago.

JOURNAL OF OUTCOME MEASUREMENT®, 2(2), 123 - 141

Controlling the Judge Variable in Grading Essay-Type Items: An Application of Rasch Analyses to the Recruitment Exam for Korean Public School Teachers

Sunhee Chae
Korea Institute of Curriculum and Evaluation

The purpose of this paper is to show how the Rasch measurement model can be used to control the effects of judge variable on the grading of essay-type items in the recruitment test for Korean teachers. Special attention is given to two aspects of judges' involvement in the grading. One is to identify a way to minimize the variation of grading due to judge severity. The other concern is to figure out a way to reduce the number of judges without threatening objectivity of ability estimates. Results from the FACETS analyses tell us not only how much grading standards vary among judges and how to adjust them but also it produces comparably reliable ability estimates with fewer judges.

Requests for reprints should be sent to Sunhee Chae, Korea Institute of Curriculum & Evaluation, Kangnam-Ku Chungdam 2-Dong 15-1, Seoul, 135-102, Korea.

To ensure high quality public school teachers in Korea, the Central Committee for Public School Teaching Credentials was formed in 1995 to reform the recruitment exam for public school teachers. The main feature of this reform was to replace multiple choice items with essay-type items. The examination system was first implemented in 1991. Prior to this all graduates with teaching certificates from teachers' colleges in national universities were allowed to teach in the public schools without going through any additional screening procedures. The introduction of the examination system in 1991 has been significant in two respects. First, it endowed every college graduate with a chance to apply for a teaching job regardless of the kind of college they attended. Second, open competition for public teachers has infused more qualified applicants into the teaching profession.

The initial examination system consisted of two steps: a multiple choice test at the national level and a general essay test and an interview at the provincial level. The first step in screening was to select a quota from all applicants taking the multiple-choice test administered at the national level. The multiple-choice test consisted of 70% subject matter and 30% general education. During the second step applicants were selected through both an essay test in general aptitude and an interview at the local level. The essay test in general aptitude consisted of the same national content, although administered separately by each province.

In an effort to ensure the high quality of school teachers through the examination system, the Central Committee for Credentials for Public School Teachers in 1995 concentrated its attention on reforming the subject matter at the national level. One method the Committee decided to adopt for the 1996 exam was to use essay-type items, instead of multiple-choice items, in all 46 subject areas (see Table 1). A preliminary mock exam was implemented in August, 1996.

It is expected that the use of essay-type items will gradually increase in the future. The underlying rationale is that essay-type items can discriminate potential teachers on the basis of more appropriate properties for teaching and interacting with students than multiple choice items, and furthermore lead pre-service teacher education in the colleges toward more desirable direction.

However, there exists some obstacles when introducing essay-type items into the recruitment exam. First of all, the grading of essay-type items presents particular challenges in terms of objectivity, since it is more likely to be affected by particular characteristics of each judge. The axi-

Table 1
46 Subject Areas

Subject Areas	Subject Areas
(1) Korean Language and Literature	(24) Mathematics
(2) Classical Chinese Character	(25) Environmental Engineering
(3) Home Economics	(26) Chemical Engineering
(4) Cloth Design	(27) Textile Engineering
(5) Social Studies	(28) Technical Engineering
(6) History	(29) Industrial Engineering
(7) Geography	(30) Machinery
(8) Moral Education	(31) Automobile
(9) Physics	(32) Electronics
(10) Chemistry	(33) Electronic Calculation
(11) Life Science	(34) Computer
(12) Earth Science	(35) Printing
(13) English	(36) Architecture
(14) Germany	(37) Mechanical Design
(15) Chinese	(38) Electromagnetic
(16) Spanish	(39) Communication
(17) Japanese	(40) Gardening
(18) Russian	(41) Tourism
(19) Physical Education	(42) Commerce
(20) Military Training	(43) Therapeutics
(21) Music	(44) Thremmatology
(22) Fine Arts	(45) Special Education
(23) Nursing	(46) General Education

omatic principle of grading is that no examinee should be advantaged or disadvantaged just because his/her responses are graded by a particular judge. If you have one judge the grading system is standardized, but not necessarily objective. This is mainly because in this situation all examinees are under the same influence of one judge, discounting intra-judge variability which is almost impossible to overcome. This type of grading, however, is unrealistic in most real tests, especially in large scale ones where judges are customarily assigned to grade only select test items. In this sense, one imperative precondition to the implementation of essay-type items is to check whether or not inter-judge variation has a significant effect upon the results of grading.

In implementing essay-type items another factor to be considered along side objectivity is economy of grading. To avoid unrealistically high costs required for the grading of all items by all judges we need to find an optimal way of partial pairings among judges so as to minimize the costs while securing the objectivity of grading. In this pairing various nesting models need to be constructed and tested to identify ways to enhance economy of grading, most presumably either by saving grading time or by reducing the number of judges, depending upon the urgency of grading. Generally, the grading of essay-type items requires more time and money than that of multiple-choice items, especially when the number of applicants and questions are substantial. This is a critical issue in large-scale testing situations, where many applicants take an exam and there is little time to score the results, as is the case in the recruitment examination for public school teachers in Korea. For example, in one province (Kyungki-Do), approximately 2,000 applicants take the exam in the subject area of Korean language and literature and results of grading must be available in a couple of weeks. Consequently, it is of utmost importance to find an efficient method for scoring essay-type items.

The purpose of this study is to show how the Rasch measurement model can be used to deal with the two aspects of judge variable in grading essay-type items. The FACETS analyses based upon the Rasch theoretical model will help to identify effects of judges upon grading by revealing the presence and degree of inter-judge variations, to adjust these variations, and to enhance economy of grading by assigning (partially paired) judges in such a way as to reduce the number of judges without sacrificing the objectivity of grading.

METHODS

Data

The data analyzed in this study was collected from a preliminary mock examination administered in August, 1996 under a cooperative venture between the Central Committee for Credentials for Public School Teachers and the Educational Evaluation Team of the Korean Educational Development Institute. The main purposes of this exam were twofold: first to check if essay-type items can be a reliable measure for screening qualified teachers, and secondly to see if the expected administrative problems can be adequately handled without seriously threatening the objectivity of grading, before the actual implementation of the recruitment exam in December, 1996. Essay-type items were employed only in some selected subject fields (Korean language and literature, English, and Mathematics) in the preliminary test. The reason for this limited selection of subject areas in the preliminary exam was that these subjects, being most popular ones, generally attracted more applicants than others. Therefore, if we can ensure the reliability of essay-type items as a screening tool in these three subjects, we expect that the remaining subjects would not cause any particular difficulties.

The three exams consisted of essay-type items and four judges were assigned to each subject area. Judges were content specialists, such as professors or teachers in their subject areas. A one-day judge training session was offered with the instructions of item constructors in each area. In the training session judges critically reviewed all the items presented by item constructors. The session continued until all judges reached a complete understanding of each item through extensive discussions. Next, the item constructors and judges participated in the construction of detailed grading criteria.

To confirm the actual applicability of these criteria, about 30 responses were randomly sampled after the mock exam and graded by judges and item producers on the basis of the criteria set during the training session. During this preliminary grading means to increase the validity and reliability of grading were additionally sought for, such that ambiguous grading criteria were discussed and rectified to be clearly understood and applied. In the actual grading, judges for each subject area gathered in the same place so that they could further discuss and reach consensus on any unanticipated responses. One divergence from this strict control of judges,

both in the training session and in the actual grading, occurred in the subject area of English. Simply due to administrative mistakes judges in this discipline did not undergo the training session and also missed a chance to be gathered in one place for grading. Though this deviation in English was not intended at first, it constituted a natural qausi-experimental situation in which one group is controlled to check the effects of judge training and control in the grading process.

The detailed information about the preliminary exam, such as the number of examinees, the number of questions and the number of judges, is presented in Table 2.

Table 2
Preliminary Examinations in Korean, English, and Mathematics

	Korean	English	Mathematics
Number of Applicants	167	156	150
Number of Items	8	12	17
Number of Judges	4	4	4

Note. Each subject area had a 2-hour session for testing.

Analysis

As outlined above, this study has two main analytical purposes; one is to determine the degree of inter-judge variations in grading essay-type items, and the other is to identify a way to reduce the number of judges needed for grading without damaging the objectivity of grading. As to the first purpose (i.e., to check the effect of judges upon grading), graded results were analyzed in two ways. First, inter-judge correlation, a traditional way of checking this effect, was computed from the grades of responses given by judges. Pearson's correlation coefficients were calculated in every pair of two judges. This method is prevalently used to confirm the reliability of judges' grading in Korea. However, the inadequacy of this traditional method to ensure the reliability of grading essay-type items requires us to employ another analysis. The second method was the Facets analyses, an extended version of the Rasch measurement model (Andrich, 1988; Rasch, 1980; Wright & Masters, 1982). The Facets model (Linacre, 1988) is an additive linear model based on a logistic transformation of the observed ratings to a logit scale. The Facets model used for this analysis was as

follows:

$$\text{Ln}[P_{nijk} / P_{nij(k-1)}] = \beta_n - \delta_i - \lambda_j - \tau_k,$$

where P_{nijk} is the probability of examinee n being rated k on item i by judge j, $P_{nij(k-1)}$ is the probability of examinee n being rated $k-1$ on item i by judge j, β_n is the ability of examinee n, δ_i is the difficulty of item i, λ_j is the severity of judge j, τ_k is the difficulty of category k relative to category $k-1$. Any variations in grading that originate from judge factor will be detected through this analysis.

As to the second purpose of this study (i.e., to identify a way to reduce the number of judges for grading), Facets analyses were applied in two ways; a benchmark analysis and a paired-judge analysis. The benchmark Facets analysis is the method whereby all participating judges evaluate all responses. This benchmark analysis was good for the production of ideal examinee ability measures since all judges graded all items for all applicants. In this situation, the existence of varying degrees of severity among judges does not necessarily cause unfair grading, because all applicants receive the same degree of severity from all the judges. A benchmark applicant ability estimate was calculated from the scores given by all four judges and all items in each subject test.

In the second Facets analysis, it was assumed that only two of the four judges were allocated to assess applicants' responses. Because all four judges had actually graded the responses of all applicants, six different combinations of judges' grades could be obtained. Six combined sets of ability measures for all applicants, in each of which only two judges actually graded responses were compared to the benchmark ability measures produced by all four judges. Any variability among ability measures in this comparison must be regarded as originating from the differences in judge severity, not from differences in applicant ability or item difficulty, since in this situation there is no change in items and applicants, but only in the judges. If all responses have similar estimated ability measures comparable to the estimated applicant measure from the benchmark analysis, then we can ensure reliability of essay-type items for a reliable screening measure and thus safely introduce them in the December exam. To confirm this point, plots comparing the benchmark ability estimates with the judge pair ability estimates were drawn. The correlation coefficients and the standardized difference scores were also calculated. Standardized difference scores were calculated from the following formula,

$$Z = (d_{i1} - d_{i2}) / (s^2_{i1} + s^2_{i2})^{1/2},$$

where $(d_{i1} - d_{i2})$ is the value of subtracting ability estimates of each set from the benchmark ability estimates, and $(s^2_{i1} + s^2_{i2})^{1/2}$ estimates the expected standard error of the difference between the two independent estimates of the examinee ability (Wright & Stone, 1979). Since each pair of ability estimates applies to the same applicant, it is expected that the two estimates are not different beyond their standard error.

RESULTS

Overall Description of Items

Table 3 shows the estimated item difficulty parameters and item separation indices for the three subject tests. Some items, like item #5 in the Korean language and literature exam and item #12 in the English exam, were off-target. Item #5 in the Korean language and literature exam requested the applicants to translate Korean characters into Chinese. This item was designed for the purpose of encouraging potential teachers to study Chinese characters during their undergraduate years, though its difficulty level was too high compared to applicants' present ability level. Item #12 in the English exam was about 'liaison,' an area not adequately covered in most undergraduate courses. Relatively greater calibrated errors occurred for math items, which may come from the fact that some items were too hard for the applicants, and too many items (17 items) were given in a limited testing time (120 minutes). The reliability of the item separation indexes are satisfactory in all three subject exams (1.00 for Korean, 1.00 for English and .99 for Math).

Inter-Judge Correlation and Judge Severity

The overall inter-judge correlation coefficient was .90 for Mathematics, .83 for Korean, and .76 for English (see Table 4). These coefficients seem to indicate satisfactory judge reliability. Various factors are believed to have contributed to this result, including the selection of judges from content specialists in the subject areas, the well-structured judge training session, the specification of detailed criteria for grading, the consensus among judges over disputed responses through discussion (except for the

Table 3
Items in Difficulty Order

Korean		
Item	Item Difficulty Logit	Standardized Error
	(most difficult)	
5	.81	.05
8	.25	.02
2	-.01	.02
1	-.03	.02
7	-.06	.02
3	-.09	.01
4	-.15	.02
6	-.72	.04
	(least difficult)	
Mean	.00	.02
SD	.40	.01
Separation 14.52	Reliability 1.00	

English		
Item	Item Difficulty Logit	Standardized Error
	(most difficult)	
12	.87	.04
7	.14	.02
6	.13	.02
10	.12	.01
8	.09	.02
9	.01	.02
11	-.04	.01
5	-.11	.02
3	-.21	.02
4	-.28	.02
2	-.32	.02
1	-.40	.02
	(least difficult)	
Mean	.00	.02
SD	.32	.01
Separation 14.68	Reliability 1.00	

Table 3
Items in Difficulty Order (continued)

Mathematics		
Item	Item Difficulty Logit	Standardized Error
	(most difficult)	
13	2.06	.23
10	.76	.09
5	.68	.08
15	.67	.09
4	.25	.06
6	.20	.05
3	.14	.05
8	.03	.04
2	-.10	.04
17	-.15	.04
14	-.18	.04
1	-.32	.03
12	-.35	.04
9	-.38	.04
18	-.44	.02
7	-.53	.04
16	-1.02	.03
11	-1.31	.04
	(least difficult)	
Mean	.00	.06
SD	.73	.05
Separation 9.68	Reliability .99	

English exam), the control of responses by their length and time consumed for their production, and so on. The slightly lower inter-judge correlation coefficient (.76) for the English exam is believed to result from the lack of consensus regarding grading criteria among the judges, since judges were not under strict control in the training session and in the actual grading due to administrative mistakes.

Table 4
Interjudge Correlations Based on Scores Graded

Korean				
Judge No.	1	2	3	4
2	.83	1.00		
4	.95	.82	1.00	
3	.79	.81	.78	1.00
English				
Judge No.	1	2	3	4
2	.81	1.00		
4	.66	.73	1.00	
3	.80	.80	.75	1.00
Mathematics				
Judge No.	1	2	3	4
2	.97	1.00		
4	.97	.99	1.00	
3	.84	.82	.82	1.00

Note. Each applicant's response was scored by four judges. Overall interjudge correlation = .90 for mathematics .83 for Korean, and .76 for English.

Table 5 presents the calibrated logit measures, standard errors, and the reliability of separation for these measures of judge severity in the three subject exams. The reliability of the judge separation indexes for the three subject exams are satisfactory overall, though it is higher for English (.99) than for both Korean (.70) and Math (.63). In the Korean exam, Judge 1 and 3 showed a high inter-judge correlation of .95, but their estimated logits were significantly different (.03 versus -.04). This finding suggests that a particular applicant was disadvantaged depending on which judge they were assigned to, even if the inter-judge correlation was sufficiently high. Inter-judge correlation is not a sufficient indicator of reliable grading. Rather, it only indicates that the pattern of a judge's grading coincides

with that of another. Despite high inter-judge correlation, judges can give significantly different scores to the same applicant simply due to differences in their severity. Consequently, varying judge severity in grading can critically threaten the objectivity of grading essay-type items. The Facets model for estimating applicants' scores in the presence of varying severity is therefore warranted.

Table 5
Judges in Severity Order

Korean		
Judge No.	Judge Severity Logit	Standardized Error
	(most severe)	
1	.03	.01
2	.01	.01
4	.01	.01
3	-.04	.01
	(least severe)	
Mean	.00	.01
SD	.02	.00
Separation 1.54	Reliability .70	
English		
Judge No.	Judge Severity Logit	Standardized Error
	(most severe)	
1	.10	.01
4	.09	.01
2	.00	.01
3	-.19	.01
	(least severe)	
Mean	.00	.01
SD	.12	.00
Separation 10.72	Reliability .99	

Table 5
Judges in Severity Order (continued)

Mathematics		
Judge No.	Judge Severity Logit	Standardized Error
	(most severe)	
4	.05	.02
1	.01	.02
2	-.02	.02
3	-.03	.02
	(least severe)	
Mean	.00	.02
SD	.03	.00
Separation 1.32	Reliability .63	

Comparison of Ability Measures Based on Benchmark vs. Paired-Judge Analysis

For each applicant, there were six sets of ability measures graded by different combinations of judges (see Table 6). Each set of ability measures can be compared to the benchmark ability measure for each subject area exam. The benchmark analyses produced ability measures for every applicant in the subject tests of Korean, English, and Mathematics with good reliability and fit statistics. The average scores for the benchmark ability measures were -.30 (.08) for the Korean exam, -.01 (.07) for the English exam and -.95 (.13) for the Mathematics exam. The average measures for the six sets of the Korean exam did not vary (-.36 to -.26) beyond the measurement error (.12). The average of the six sets for the mathematics exam also did not vary (-1.0 to -.93) beyond the measurement error (.19). Contrary to these findings, the average score of estimated measures in the English exam was more variant (-.11 to .10 with the measurement error of .10) compared to the other two exams, albeit not seriously.

The correlation coefficients and the standardized z-scores also confirmed these findings. Figure 1 shows that the correlation coefficient between the ability measures scored by judges 1 and 2 and the benchmark

Table 6
Applicant Ability Measures Graded by Six Different Pairs of Judges

Judge Combination #s	Korean (N = 167)		English (N = 156)		Math (N = 150)	
	Logit	Error	Logit	Error	Logit	Error
1 & 2	-.36	.13	-.11	-.09	-1.0	.18
1 & 3	-.29	.12	-.03	.10	-1.0	.18
1 & 4	-.29	.12	-.04	.10	-.97	.19
2 & 3	-.27	.12	.00	.10	-.97	.18
2 & 4	-.28	.12	-.10	.10	-.93	.19
3 & 4	-.26	.11	.10	.10	-.93	.19
Mean	-.29	.12	-.03	.10	-.97	.19
SD	-.03	.01	.07	.00	.03	.01
Average Benchmark Ability Estimates	-.30	.08	-.01	.07	-.95	.13

ability measures for the Korean test is .98. The rest of the figures comparing the ability measures graded by all paired judges to the benchmark ability measures for all three subject tests show results similar to the one identified in Figure 1. The correlation coefficients for all the comparisons ranged from .95 to .98. Based on the standardized z-scores, no examinee obtained significantly different ability measures from their benchmark measures at the 95% significance level for the Korean and Mathematics exams, even though they were graded by different pairs of judges. However, in the English exam, some portion of the applicants appeared to get somewhat varying ability measures especially when judge #1 was paired with #2, and judge #3 with #4 (see Table 7). This result can be explained by the fact that the four judges in the English exam had little chance to reach consensus in the grading process because they graded separately without being assembled in one place. In addition, judges commented that some of the items were selected from very unpopular areas where even the judges, being unfamiliar with them, had difficulty in grading them. This finding can be confirmed by different judge severity measures among the subject

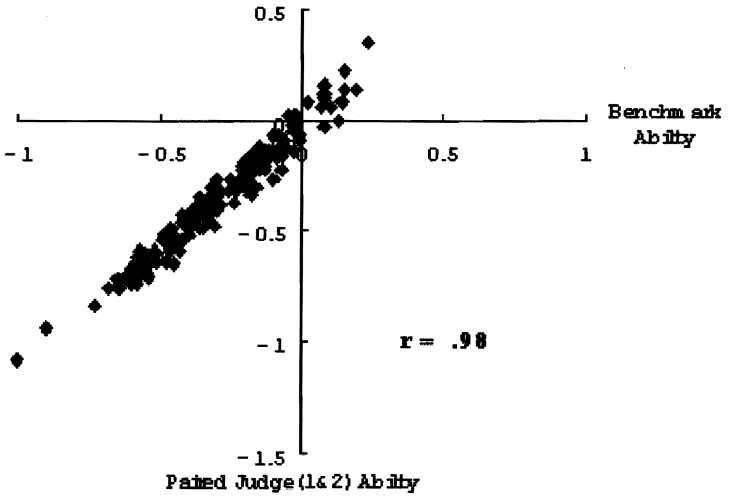


FIGURE 1 Ability Measures Based on Benchmark Analysis vs. Paired Judge(1&2) Analysis- Korean.

areas, as shown in Table 5. Among the three subjects, the judges for the English exam manifested a highest variability (-.19 to .10) in their severity measures, compared with those of the Korean exam (-.04 to .03) or of the Mathematics exam (-.03 to .05). This informs us that we can detect statistically significant variability of examinee ability measures in paired grading from the judge severity measures estimated by Facets analysis.

CONCLUSION AND DISCUSSION

So far, this paper has attempted to find ways to control effects of judge variable upon grading of essay-type items by utilizing the data obtained from a mock recruitment exam of Korean teachers. Two aspects of judge

Table 7
Standardized Difference Scores in Ability Comparisons in the English Exam

Six Comparisons	Benchmark vs. Judge 1&2	Benchmark vs. Judge 1&3	Benchmark vs. Judge 1&4	Benchmark vs. Judge 2&3	Benchmark vs. Judge 2&4	Benchmark vs. Judge 3&4
Z-scores Bigger than 1.96 in Ability Comparisons of Benchmark vs. Paired-Judge Analysis	3.52 (146)	2.11 (9)	2.08 (25)	2.23 (31)	2.55 (104)	-1.99 (55)
	3.05 (147)	2.09 (107)	1.98 (88)	2.23 (76)	2.40 (94)	-2.00 (98)
	2.87 (93)		-2.38 (8)	-2.02 (148)	2.23 (146)	-2.05 (153)
	2.82 (18)				2.06 (39)	-2.07 (123)
	2.68 (76)				2.06 (126)	-2.10 (95)
	2.63 (99)				2.01 (85)	-2.16 (152)
	2.59 (39)				1.98 (56)	-2.22 (4)
	2.50 (4)					-2.22 (39)
	2.50 (50)					-2.38 (64)
	2.24 (156)					-2.40 (50)
	2.23 (111)					-2.45 (44)
	2.22 (100)					-2.46 (49)
	2.14 (79)					-2.60 (27)
	2.13 (20)					-2.61 (137)
	2.13 (84)					-2.76 (28)
	2.13 (155)					-2.83 (147)
	2.03(104)					-2.84 (12)
	1.98 (28)					-3.00 (93)
	1.96 (25)					-3.00 (146)
						-3.31 (18)
The number of Z-scores bigger than 1.96	19 (12.2%)	2 (1.3%)	3 (1.9%)	3 (1.9%)	7 (4.5%)	20 (12.8%)

Note. Examinee id numbers are in the parenthesis.

variable were given special attentions. One is to figure out a way to verify the presence and degree, if any, of judge severity upon the grading. For this purpose, the traditional method of inter-judge correlation coefficients were compared with the results of Facets analyses. Through this comparison we learn that the traditional method was insufficient to accurately detect disparity among judges in grading essay-type items. The main reason was that correlation coefficients only show whether or not the grading patterns among the judges coincide. If the judges grade responses in the same direction (that is, if the responses are graded in the same ranking order among the judges), correlation coefficients turn out to be very high even if the actual grade each response receives differs significantly. This testifies to the possibility that a particular applicant can be disadvantaged by the judge severity even if correlation coefficients of grading are very high.

The Facets model is very efficient to overcome this problem as it incorporates the judge variable in its actual analyses. It not only enables us to detect differences in the degree of judge severity which go unnoticed in the traditional method, but more importantly produces examinee parameters that have adjusted the judge severity and thus is devoid of subjective vaccillation. These parameters are a reliable measure upon which we can make solid decision-makings about applicants' ability. Obtaining these results, we are safe to introduce essay-type items both to the actual recruitment test of Korean teachers and more generally to other types of performance assessments.

The other aspect of judge variable this study is concerned with is to identify a way to reduce the cost of grading essay-type items. Unlike multiple choice items to which computerized grading is applied, essay-type items demand enormous costs in their grading if we are not to sacrifice the objectivity. Considering the positive educational functionality essay-type items have, it is quite imperative to find a way to enhance the economy of grading them if they are to be widely implemented even in non(or less)-affluent countries. This study approached the problem of economy mainly by finding a way to reduce the number of judges participating in grading without threatening objectivity.

One solution suggested in this study is to compare ability measures of applicants based on the benchmark grading with those of paired judge grading. In the former, an ideal situation of grading was assumed in which all judges graded all responses. And in the latter, the judges were assigned in such a way that two judges were to comprise each pair in rotation, all in

all constituting six pairs. Comparison findings from the benchmark grading and paired judge grading indicate that each paired-judge grading produced results that are statistically equivalent at the 95% significance level, and that correlation coefficients were higher than .95 in all three subjects. After all, examinee parameters obtained from the Facets analyses suggest a way to enhance the economy of grading by reducing the number of judges, since they confirm that even employing fewer judges in grading, we can achieve fairly objective grading. This implies that the paired grading, only if well-designed according to the Facets model, will lead a way to reduce the costs of grading essay-type items.

The divergence in the grading of English from the general pattern, however, proposes an important caveat. In the English exam we find a little higher examinee severity measure, albeit statistically insignificant. This reminds us of the inordinate importance of strictly controlling judges both in the training sessions and in the actual grading, such as selection of qualified judges, sufficient time for judge training, the specification of detailed and unambiguous criteria for grading, and the consensus among judges through discussions. Without these precautions, all other results could be endangered.

The implementation of essay-type items in the actual recruitment exam of Korean teachers in December 1996 was significantly encouraged by the results of this preliminary study. Furthermore, this study, when complemented by other similar studies, could provide a good empirical ground upon which various kinds of performance assessment can be introduced in diverse Korean educational settings, as it suggests means to deal with some problems expected from the introduction of essay-type items in large-scale testing.

ACKNOWLEDGEMENTS

The data used for this paper were collected by the research team (I. Sung, Y. Kim, S. Baek, & S. Chae) of the Korean Educational Development Institute. I thank them for allowing me to use this valuable data for my research.

REFERENCES

Andrich, D. (1988). *Rasch models for measurement.* Newbury Park: Sage Publications.

Linacre, J. M. (1988). *FACETS, a computer program for analysis of examinations with multiple FACETS.* Chicago: MESA Press.

Rasch, G. (1980). Probabilistic models for some intelligence and achievement tests (rev. ed.). Chicago: University of Chicago Press. (Original work published 1960).

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

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

JOURNAL OF OUTCOME MEASUREMENT®, 2(2), 142 - 158

Team Assessment Utilizing a Many-Facet Rasch Model

Jeff M. Allen
and
Randall E. Schumacker
University of North Texas

As organizations begin to implement work teams, their assessment will ultimately reflect compensation strategies that move away from individual assessment. This will involve not only using multiple raters, but also the use of multiple criteria. Team assessment using multiple raters and multiple criteria is therefore necessitated; however, this can produce differences in ratings due to the leniency or severity of the individual team raters. This study analyzed the ratings of individual members on 31 different teams across 12 different criteria of team performance. Utilizing the many-facet Rasch model, statistical differences between the teams and 12 criteria were calculated.

Requests for reprints should be sent to Jeff M. Allen, University of North Texas, P. O. Box 311337, Denton, TX 76203.

With the advent of teams in the workplace, there is a need for better and more valid assessment processes, especially for work teams making multiple types of decisions, including compensation. Most of the employee compensation systems that exist in companies today are inadequate to effectively reward the success of a team-oriented work force. Because there are few measures of team development and performance, compensation systems often rely on an individual supervisor to measure the individual contributions to a team in determining an employee's year-end salary increase. This is an ineffective compensation system when one takes into account the new work practices that organizations have fought to gain which focus on team success.

Seventy-three percent of organizations were using group or team compensation in 1993, as compared to 59% in 1990 (Bassi, Benson, & Cheney, 1996). The work team concept has introduced a new challenge for human resource development personnel. This challenge opens many new opportunities in the employee evaluation and training fields. Organizations are demanding that employees recognize the organization as a quality system. This not only requires that departments and supervisors move in the same strategic direction as an organization, but also that individual employees interact with others in the organization to help the organization succeed. As an individual's job becomes more complicated and more reliant on others in the organization, an organization's compensation system must comprehensively evaluate an individual employee's contribution to the organization in that context. Today, the trend is moving toward team assessment instead of individual assessment. "90 percent of Fortune 1000 companies now use some form of multi-source assessment" (Industrial Report, 1996, p. 24).

Training magazine reported that 36% of industrial teams are responsible for their own performance appraisal (Industrial Report, 1996). One of the ways to approach the assessment of a team is to use multiple criteria and multiple raters. The supervisors in the organization cannot alone, in this new team-oriented approach, comprehensively evaluate their employees or teams of employees on a single criterion. An employee's team-oriented job sometimes crosses many different organizational boundaries. In this situation an employee interacts with many people inside and outside of the organization. In some environments, the supervisors must rely on other raters (other supervisors, team leaders, or an employee's peers) to contribute to the comprehensive assessment of an employee's performance. Bracken (1994) has noted that multi-rater systems are on the rise, and

more and more organizations design and implement processes in which employees are rated by some combination of managers, supervisors, peers, direct reports, and even customers. As the employee evaluation and training fields move toward more complex evaluation issues, multiple criteria and multi-rater assessments will become even more popular, useful, and relevant. Consequently, there is a need to better understand team assessment using multiple criteria, and especially multi-rater assessment, since team evaluation ratings are typically given by members of the team.

The purpose of this study was to analyze team performance on multiple criteria, given that the ratings were from team members. Utilizing the many-facet Rasch model, we demonstrate that multiple criteria differences in team ratings can be utilized to better assess the "meaning" of team performance.

METHODOLOGY

Instrumentation

A pilot study (not reported here) was initially conducted to assess the content validity and reliability of items that were developed for the instrument. Manufacturing organizations were selected in which five teams were identified, with from 4 to 23 team members on each team, for a total of 58 team members. The identified teams were active in the organization and were involved in various manufacturing and service operations. The objective of the pilot study was to determine whether increasing or decreasing either the number of items per scale or the number of team members on each team would affect the reliability of the assessment instrument. The pilot study yielded acceptable initial reliability estimates for the scales in the assessment instrument. Results of the pilot study were encouraging; thus, no changes in the scales on the instrument were needed.

The team assessment instrument contained 70 content-valid items across 12 criteria. The 12 criteria and number of items per criteria in parentheses were: Data-Based Decisions(4), Ground Rules(2), Team Comfort(3), Team Difference(3), Cooperation(9), Resources(3), Satisfaction(10), Problem Solving(4), Quality(10), Open Communication(5), Planning(13), and Decision Making (4). The items used in each of the criteria are listed in the Appendix. The items on the instruments used a Likert scale from 1 to 5, where 1= *strongly disagree*; 2= *disagree;* 3= *neutral*; 4= *agree*; and 5= *strongly agree*. For our sample of participants, the Cronbach alpha internal consistency reliabilities for the 12 criteria (scales)

and the number of items for each criteria are reported in Table 2. With the exception of the Resource scale, the other scales had acceptable levels of reliability, especially given the number of items (Given space limitations we did not include inter-item correlation matrices for the scales).

Participants

The target sample consisted of 308 team members from a population of 372 available personnel. The selected industrial organizations comprised military, electronic, and apparel operations located in central Pennsylvania. The first author and a representative from each participating organization identified the work teams for the study. There were 31 work teams identified, with from 2 to 32 members on each. The number of team members on each team is identified in Table 1. The rationale for selecting the sample was to diversify the number of teams and team members, but not invoke a diverse number of organizations. All teams participating in this tudy were active in their organizations, and all the teams had existed for 2 weeks to 4 years. Sixty-four instruments were unusable (incomplete data, or uninterpretable) from the 372 available personnel, yielding the resulting sample of 308.

Procedures

The assessment instrument was administered to each team during its scheduled team meetings. During the administration of the instrument, instructions were explained and individual team members rated their team on each of the 12 criteria. Team members were allowed to ask the administrator clarifying questions to better understand the relationship of a general question to their specific team. Each team member rated their own team on all items, but did not rate individual team members. The typical adminstration time for the instrument was approximately 35 minutes.

Data from each team member were entered into a data matrix (308 team members x 12 criteria ratings). These data were further coded to reflect the 31 individual teams. Each team comprised a different group of team members, so making comparisons between the 31 teams on a specific criterion reflects average differences in each group of team members on that criterion. When comparing a specific team across the 12 criteria, a relative comparison among the criteria is possible to reflect which of the criteria a given team rated the highest. A team could therefore compare

itself across the 12 criteria or examine how its ratings compared to those of another group of team members. Obviously, team members could introduce a rating bias of their team.

From a design perspective, raw scores were obtained from team members within each team on 12 criteria. The many-facet Rasch model converted the ratings of team members on the criteria into logits. The extent to which teams or the criteria differed was found by examining the logit measure differences (main effects). The chi-square statistic reported by the Facets computer program tested whether the levels of these main effects (teams or criteria) were significantly different.

How the ratings were obtained (design), as well as the scoring and analysis methods used, has a significant impact on the decision outcome. The influence of each facet (teams or criteria) on the rating score should be observed. The basis for validity is the meaning assigned to the scores (Messick, 1995); therefore, it is helpful to understand as fully as possible how the score is derived. For example, one team may be rating the criteria with more *severe* team members, while another team may be rating the criteria with more *lenient* team members. The rating scores obtained would therefore have a very different "meaning," depending upon the leniency or severity of the team members (raters). The many-facet Rasch analysis adjusts the raw score ratings for this leniency-severity rater factor, which is not considered when simply averaging raw score ratings.

Analysis

In the rating scale model for this study, the three-facet Rasch model was written as follows: $log\ [P_{nijx} / P_{nij(x-1)}] = B_n - D_i - C_j - F_{ix}$. The terms are defined as follows:

P_{nijx} = probability of team n being rated x on criteria i by rater j.
$P_{nij(x-1)}$ = probability of team n being rated x-1 on criteria i by rater j.
B_n = performance of team n
D_i = difficulty of criteria i
C_j = effect of rater j
F_{ix} = difficulty of observed category x relative to category x-1 (step difficulty)

Note that the performance of a team on the criteria is adjusted for the effect of rater j. The many-facet Rasch model permits an adjustment to the team performance ratings, based upon the particular group of raters (team members) who rated it and what criteria. This is a distinct advantage of the

many-facet Rasch model in conducting team assessment over simply interpreting the average raw score ratings.

The raw rating scores were input into a Facform program that then produced a comma-separated data file and FACETS program suitable for analysis (Linacre, 1994). The FACETS program outputed calibrated logit values for each element of a facet (level of the variable). In addition, a chi-square test examined the similarity among the facet elements: a "fixed" effects chi-square test was computed to test whether the L measures were statistically equivalent to one common "fixed" effect apart from measurement error. If $p < .05$, then L facet element measures are statistically different.

FINDINGS

Table 1 indicates the observed score (sum of ratings), observed count (frequency of ratings), average rating (observed score divided by observed count), logit measure, and the standard error of the logit measure for each of the 31 teams. The fixed chi-square value of 648.3, $df = 30$, $p = .001$ indicates that the teams are significantly different in performance on the 12 criteria, based on the team member ratings. The logit measures ranged from .24 to 2.98. The higher logit measures imply a higher team performance measure when compared to the other teams. This logit scale also permits a linear equal interval interpretation of just how much each team differs in its ratings. Notice that the internal consistency of the ratings for the teams was .93.

Table 2 indicates the same type of information as in Table 1, but for the 12 criteria that were rated by the individual team members on a team. The fixed chi-square value of 233.6, $df = 11$, p =.001 indicates that the scales were significantly different in how they were rated. The logit measures ranged from -.65 to +.68, with the negative logit meaures indicating higher average scale ratings. Data-Based Decisions, Ground Rules, Team Comfort, Team Differences, and Cooperation were rated lower than Decision Making, Open Communication, and Planning. Once again, the magnitude of difference between the criteria is depicted on an equal interval linear scale. The criteria ratings by individual team members had an internal consistency reliability of .95.

The 308 raters were also significantly different in their ratings ($\chi^2_{fixed} = 1537$, $df = 307$, p=.0001). The logit rating measures ranged from -4.11 to 2.37 with negative logit measures indicating higher ratings.

Table 1
Team Comparisons

Obsvd Score	Obsvd Count	Obsvd Average	Fair Average	Logit Measure	Model S.E.	Infit MnSq	Infit Std	Outfit Mn Sq	Outfit Std	Team Num.	No. of Members
346	84	4.1	4.1	2.98	.20	0.6	-3	0.6	-3	8	7
337	84	4.0	4.0	2.71	.20	0.9	0	0.9	0	11	7
287	72	4.0	4.0	2.58	.22	1.2	1	1.2	1	14	6
471	120	3.9	3.9	2.34	.16	1.2	1	1.1	1	24	10
227	60	3.8	3.9	2.23	.24	1.0	0	0.9	0	4	5
185	48	3.9	3.9	2.19	.25	1.9	3	2.0	3	16	4
92	24	3.8	3.8	2.04	.36	1.1	0	1.1	0	18	2
320	84	3.8	3.8	2.00	.19	0.8	-1	0.8	-1	9	7
411	108	3.8	3.8	1.98	.17	0.9	0	0.9	0	23	9
273	72	3.8	3.8	1.93	.21	1.1	0	1.1	0	15	6
634	168	3.8	3.8	1.88	.14	0.9	0	0.9	0	28	14
450	120	3.8	3.8	1.80	.16	0.7	-2	0.7	-2	6	10
358	96	3.7	3.8	1.79	.18	0.7	-2	0.7	-1	5	8
180	48	3.8	3.8	1.79	.25	1.0	0	1.0	0	17	4
448	120	3.7	3.8	1.77	.16	0.9	-1	0.9	-1	19	10
265	72	3.7	3.7	1.68	.20	1.0	0	1.0	0	7	6
219	60	3.7	3.7	1.50	.22	0.6	-2	0.6	-2	10	5
347	96	3.6	3.6	1.40	.18	0.7	-2	0.6	-2	20	8
434	120	3.6	3.6	1.39	.16	1.3	2	1.4	2	25	10
468	132	3.5	3.6	1.21	.15	1.3	2	1.3	2	29	11
679	192	3.5	3.6	1.17	.12	1.0	0	1.1	0	3	16
377	108	3.5	3.5	1.09	.16	1.0	0	0.9	0	21	9
291	84	3.5	3.5	.98	.18	0.9	0	0.9	0	12	7
1062	312	3.4	3.4	.78	.10	0.8	-1	0.9	-1	30	26
392	276	3.4	3.4	.75	.10	1.4	4	1.4	4	1	23
242	72	3.4	3.4	.63	.20	1.1	0	1.1	0	13	6
401	120	3.3	3.3	.58	.15	1.4	2	1.4	2	22	10
355	108	3.3	3.3	.47	.16	1.2	1	1.2	1	26	9
1258	384	3.3	3.3	.43	.08	0.8	-2	0.8	-2	31	32
271	84	3.2	3.2	.28	.18	1.0	0	1.0	0	27	7
537	168	3.2	3.2	.24	.13	0.0	-1	0.9	-1	2	14

Note. Fixed (all same) chi-square: 648.3 d.f.: 30 significance: .00
Random (normal) chi-square: 30.2 d.f.: 29 significance: .41
Reliability = .93.

Table 2
Criteria Comparisons

Obsvd Score	Obsvd Count	Obsvd Average	Fair Average	Measure	Model S.E.	Infit MnSq	Infit Std	Outfit Mn Sq	Outfit Std	Relia-bility	No. of Items	Criteria
1024	308	3.3	2.9	.68	.09	0.9	0	1.0	0	.59	4	Data Based Decisions
1034	308	3.4	2.9	.59	.10	2.1	9	2.2	9	.52	2	Ground Rules
1048	308	3.4	3.0	.46	.10	1.3	3	1.3	2	.48	3	Team Comfort
1071	308	3.5	3.0	.25	.10	1.1	0	1.1	0	.55	3	Team Differences
1075	308	3.5	3.1	.21	.10	0.6	-6	0.6	-6	.73	9	Cooperation
1088	308	3.5	3.1	.09	.10	0.8	-2	0.8	-2	.21	3	Resources
1097	308	3.6	3.1	.00	.10	0.9	0	0.9	-1	.75	10	Satisfaction
1131	308	3.7	3.3	-.33	.10	1.1	1	1.1	1	.60	4	Problem Solving
1136	308	3.7	3.3	-.37	.10	0.8	-2	0.8	-2	.80	10	Quality
1141	308	3.7	3.3	-.42	.10	0.5	-6	0.5	-6	.81	13	Planning
1148	308	3.7	3.3	-.49	.10	0.8	-2	0.8	-2	.63	5	Open Communication
1164	308	3.8	3.4	-.65	.10	1.0	0	1.0	0	.66	4	Decision Making

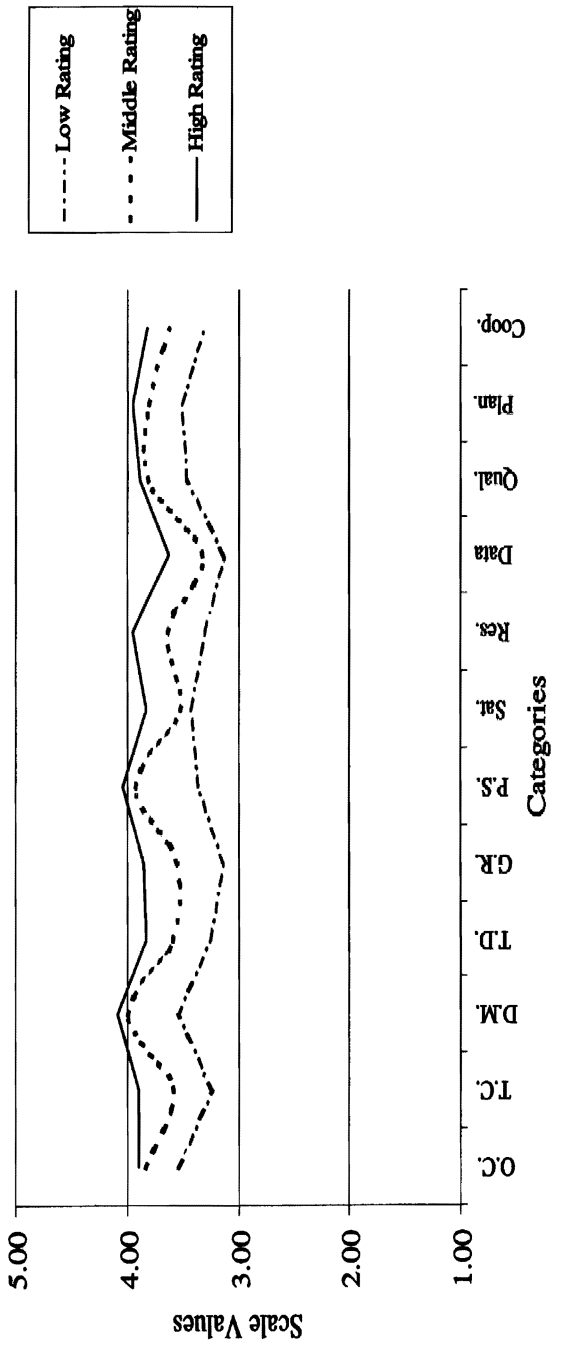
Note. Fixed (all same) chi-square: 233.6 d.f.: 11 significance: .00
Random (normal) chi-square: 11.0 d.f.: 10 significance: .361
Reliability = .95.

Raters had a reliability of .81 (A separate table was not included due to space limitations).

Figure 1 provides a comparison of the means of the 31 teams across the 12 criteria. Since a graphic comparison of the logit scores of 31 teams across 12 criteria is visually difficult, the 31 teams were recorded into three groups based on the range of logit measures. The 31 teams were evenly divided into three groups (Low rating - 11 teams, Middle rating - 10 teams, and High rating - 10 teams). Figure 1 provides an easy visual distinction between teams across multiple criteria. The figure also visually indicates that the team ratings were negatively skewed, with averages between 3.13 and 4.09.

Figure 2 provides a diagram for comparing the distribution of teams, criteria, and raters. This figure visually indicates that the team ratings were negatively skewed, implying that team members rated their respective team consistently high on performance. The diagram further indicates that Data-Based Decisions and Ground Rules, Cooperation and Resources, Problem Solving and Quality, and Open Communication and Planning were rated equivalent in importance. The logit measures can be rescaled using a suitable mean and standard deviation to aid interpretation (mean +/- logit * standard deviation). Although ratings were consistent (reliable), the narrow range of logit values in Table 2 make discrimination (validity) between the 12 criteria difficult.

The number of scale points used for rating the 12 criteria may need to be reduced. *Strongly disagree* was indicated 20 times in contrast to *Strongly agree* which was selected 344 times (Table 3). The percentages indicated that 39% of the team members selected a 3 and 45% of the team members selected a 4. This frequency count across the scale points clearly shows the tendency of team members to rate their teams above average on performance across the 12 criteria. More importantly, raters are not using all five scale points, so the number of categories used by the raters to rate the criteria could be reduced from five to two, categories 1, 2 and 3 = 1 or disagree, 4 and 5 = 2 or agree). Also, the neutral scale point could be dropped. Obviously, these concerns indicate that the measurement scale could be dichtomized without loss of meaning in future use of this instrument.


Low Rating
Middle Rating
High Rating
Scale Values
5.00
4.00
3.00
2.00
1.00
Categories
O.C.
T.C.
D.M.
T.D.
G.R.
P.S.
Sat.
Res.
Data
Qual.
Plan.
Coop.

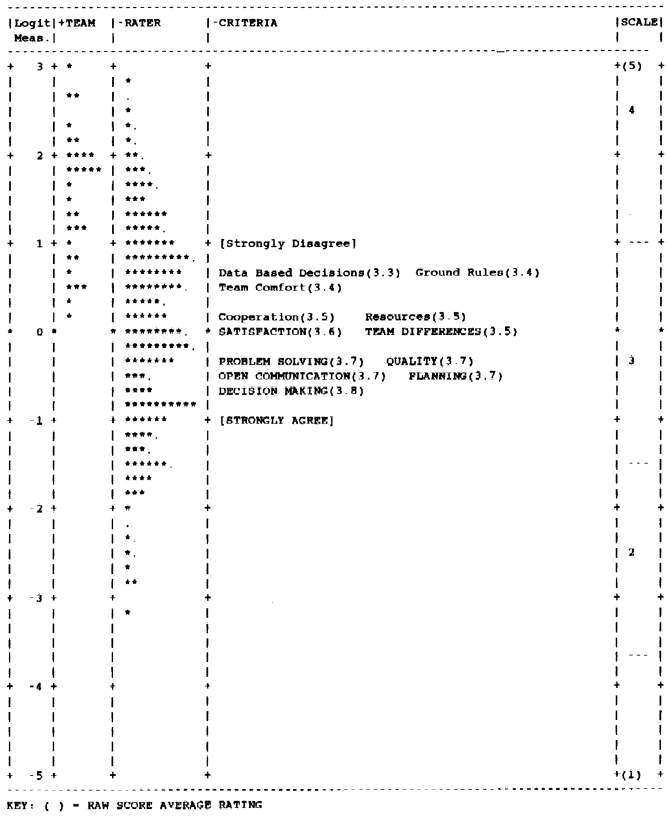


FIGURE 2 Facet Comparisons.

Table 3
Rating Scale Statistics

DATA Category Counts Cum.				Avge Meas	OUTFIT	Step Calibrations		Expectation Measure at		Most Profitable	Thurstone Threshold	Cat Peak	Obsd-Expd Diagnostic
Score	Used	%	%	Diff	MnSq	Measure	S.E.	Category	-0.5	from	at	Prob	Residual
1	20	1%	1%	-.76	1.3			(-4.52)		low	low	100%	
2	243	7%	7%	-.37	1.1	-3.31	.23	-2.57	-3.67	-3.31	-3.48	51%	-1.0
3	1425	39%	46%	.43	.9	-1.78	.07	-.36	-1.56	-1.78	-1.67	66%	-1.0
4	1664	45%	91%	1.84	.9	.96	.04	2.55	1.00	.96	.97	71%	1.9
5	344	9%	100%	3.31	1.0	4.12	.07	5.23	4.23	4.12	4.15	100%	
								Mean		Modal	Median		

DISCUSSION

The many-facet Rasch model was used to indicate differences in teams on 12 criteria based upon ratings by team members. A separate calibration was possible to compare differences among the teams and similarly to compare differences among the 12 criteria used to make up the team ratings. A comparison of the teams was therefore indicated while taking into consideration the importance of the 12 criteria. Significant differences among the teams were found, as well as significant differences among the 12 criteria used. The ability to separately calibrate and compare levels of one variable (facet), taking into consideration the levels of another variable (facet) is a unique feature of the many-facet Rasch model. Although this study only investigated the main effects or differences in the levels of the facets, interaction effects between the facets is also possible to further determine whether certain teams were rated higher or lower on any specific criteria.

Implications for Researchers and Practitioners

As organizations begin to implement work teams, their assessment will ultimately reflect compensation strategies that move away from individual assessment. However, in a team-based workforce, issues arise concerning team-based compensation and evaluation. In reality, teams are given one evaluation, typically at the end of a project. This only gives the team a post-hoc evaluation of their success or failure.

Teams are a difficult entity to understand for most organizations. How do you compensate a team for its success or failure on a project when "a team" does not exist in reality? Only a group of individual team members really exist, consequently organizations require the team members and supervisors to evaluate "the team". Training professionals are required to provide training for "the team" and human resource professionals are required to provide compensations for "the team". One of the first steps in evaluating teams in an organization is to have an understanding that "the team" is being evaluated, yet individual team members are being compensated for "the teams" success.

Examples of evaluation instruments that offer organizations a variety of information on team success are vital. In 1996, 90% of Fortune 1000 companies were using some form of multi-source assessment. Also, 36% of organizational teams were responsible for their own performance ap-

praisal (Industrial Report, 1996). The reality of team evaluation bias and the necessity of team comparisons within an organization, requires training and human resource development professionals to search for new methods to obtain non-bias information. The many-facet Rasch analysis allows us to use a multiple criteria instrument for evaluation of teams in the workplace.

The many-facet Rasch model analysis provides adjustments in raw score ratings to compensate for lenient or severe raters. It is assumed that the team members are giving an honest and fair opinion of their teams' performance. Severity-leniency of raters (team members) is seldom considered in most team evaluations. The validity of the ratings is critical when compensation decisions are made based on these rating scores. The ease in which the rating scores can be converted for Rasch analysis makes it possible, in an organizational environment, for training and human resource professionals to use the many-facet Rasch to analyze and compare teams.

The results of the many-facet Rasch analysis provides human resource development professionals a wealth of information that is not found in traditional compensation models. As an example, Table 1 provides a human resource professional critical information in the comparison of each team within an organization, without rater bias. In addition, Table 2 provides a training professional critical information for the planning of organizational training programs. Organizational training programs on data-based decisions, ground rules, and other facets would provide the team information needed to improve their team performance, and thus increase the compensation that they receive. If teams are to survive into the next decade, researchers must provide training and human resource personnel usable analysis tools to provide fair compensation systems for work teams.

REFERENCES

Bassi, L., Benson, G., & Cheney, S. (1996). The top ten trends. *Training and Development, 50*(11), 28-42.

Braken, D.W. (1994). Straight talk about multirater feedback. *Training and Development, 48*(9), 44-51.

Industrial report 1996:Trends. (1996). *Training, 33*(10), 67-71.

Linacre, M. J. (1994). *A user's guide to FACETS.* Chicago: MESA Press.

Messick, S. (1995). Standards of validity and the validity of standards in performance asssessment. *Educational Measurement: Issues and Practice,14*(4), 5-8.

APPENDIX

INSTRUMENT ITEM LISTING

Open Communication
Team members explore rather than debate each speaker's ideas.
Team members exchange ideas in many different ways. (ex. brainstorming, discussions, presentations, etc.)
Team members listen well during meetings.
Team members allow a speaker to finish his or her statement before replying.
Individual team members seek information and opinions from other team members.

Team Comfort
Individual team members do not feel comfortable beginning a discussion.
The team does not "leave out" individual team members.
Team members do not allow all members to share ideas.

Decision Making
Team members agree on the team's goals.
Team members compromise in the decision making process.
The team discusses how to make decisions. (ex. polls, votes, or concensus)
The team decides important issues by consensus.

Team Differences
Team members do not resolve differences.
Team members try to ease tensions between members.
Team members try to find the root cause of group behavior problems.

Ground Rules
The team members do not have ground rules about group behavior.
The team has openly discussed group behavior ground rules.

Problem Solving
Team members understand each member's individual roles.
Individual team members are not satisfied with the team's progress toward their goal.
The team size is not appropriate to effectively complete project goals.
The team members do not work together to achieve the team's goals.

Satisfation

The team's project activities are not important to their customers.
The team effectively delivers their products or services to their customer(s).
The team effectively implements feedback from their supervisor(s).
The team effectively implements feedback from their customer(s).
The team follows up with product or service satisfaction of customers.
The team obtains customer support for project goals.
The team obtains supervisor support for project goals.
The team effectively communicates with its customer(s).
The team effectively communicates with its supervisor(s).
The team has not been rewarded for effective performance.

Resources

The team effectively gathers resources needed to complete its project.
Individual team members seek information and opinions from outside sources.
The team does not have access to resources needed to complete project goals.

Data Based Decisions

The team uses basic statistical tools to improve their final product.
The team uses data as a base for its decisions.
The team uses basic statistical tools to make meeting decisions.
Team discussions do not stay on the current meeting topic.

Quality

The team stays within project time frames.
The team completes projects on time.
The team reliably completes projects goals.
Team goals are in agreement with organizational goals.
The team does not meet project goals.
The team's actions lead to a positive change in the company's performance measures.
The team knows which documents and reference materials are available to guide their project progress.
The team refers to written project documents to guide its project direction.
The team has documents that describe project steps.
The team produces high quality work.

Planning

The team frequently reviews its progress toward project goals.
The team continuously improves project plans to more effectively produce their products or services.
The team establishes realistic time frames for the completion of projects.
The mixture of team knowledge and skills is not appropriate for the project requirements.
The team has clearly established goals.
The team's goals are not specific enough to have achievable results.
Team members clearly understand the steps needed to reach the team's goal.
Team members do not know the purpose of team meetings.
The team has no improvement plan.
The team seeks permanent solutions to problems rather than quick fixes.
The team does not effectively plan projects.
The team does not have the relevant knowledge to successfully complete project goals.
The team effectively implements project plans.

Cooperation

Team members demonstrate poor communication skills.
The team has formally designated roles for each member.
The team does not use each individual's talents.
Individual team members speak clearly and directly to the issues.
Individual team members do not suggest procedures for reaching team goals.
Team members clarify and elaborate on other member's ideas.
Team members summarize ideas.
The team effectively communicates with other teams in the organization.
Individual team members complete assigned task.

JOURNAL OF OUTCOME MEASUREMENT®, 2(2), 159 - 167

Round-Off Error, Blind Faith, and the Powers That Be: A Caution on Numerical Error in Coefficients for Polynomial Curves Fit to Psychophysical Data

Vincent J. Samar
National Technical Institute for the Deaf
Rochester Institute of Technology
Otolaryngology Division, University of Rochester Medical Center

Carol Lee De Filippo
National Technical Institute for the Deaf
Rochester Institute of Technology

Graphing and statistics software often permits users to fit polynomial curves, like a parabola or sigmoid, to scatter plots of psychophysical data points. These programs typically calculate the curve using double- or extended-precision numerical algorithms and display the resulting curve overlaid graphically on the scatter plot, but they may simultaneously display the equation that generates that curve with numerical coefficients that have been rounded off to only a few decimal places. If this equation is used for experimental or clinical applications, the round-off error, especially on coefficients for the higher powers, can produce anomalous findings due to systematic and extreme distortions of the fitted curve, even artifactually reversing the algebraic sign of the true slope of the fitted curve at particular data points. Care must be exercised in setting round-off criteria for coefficients of polynomial terms in curve-fit equations to avoid nonsensical measurement and prediction.

Requests for reprints should be sent to Vincent J. Samar, Department of Applied Language and Cognition Research, National Technical Institute for the Deaf, Rochester Institute of Technology, Rochester, NY 14623.

When in the course of curvilinear events it becomes necessary to fit a second- or third-order polynomial function to a scatter plot of data observations such as the psychophysical data in Figure 1, most of us these days turn to our trusty graphing or statistical software. These are software applications such as CA-Cricket Graph III 1.5 (1993) that run on our office and laboratory PCs with little or no technical guidance from us. With a "pull down" here and a "double click" there, these packages will wink up a lovely least squares line that courses gracefully through the tousle of data points in the scatter plot, such as curve 1 in Figure 1.

This curve is the sculpture in the stone. The application software is the sculptor that uncovers it, and, more often than not, we are merely patrons of the sculptor's work, appreciative of the curve's form but having largely implicit, some would say blind, faith in the procedural details of its production. Always, we hope that such parabolas and sigmoids reflect some perfect relationship that underlies the error-ridden cloud of direct observations we started with. We trust that, like Plato's perfect triangles, an essential form has been recollected from the pale reality of the scatter.

One obvious application of such curve fitting is to obtain measurements on psychophysical functions in studies that test experimental hypotheses, predict clinical outcomes, or predict the performances of new individuals. This is done by estimating some parameter of the curve, for example its value or its slope at some point, which is precisely what one of us (CD) had in mind when she collected the data set in Figure 1, along with 73 other data sets just like it. The data in Figure 1 are part of a cross-modality matching study whose purpose is to establish the psychophysical function between visual line length (ordinate) and pure tone sound intensity (abscissa) in 25 individuals with mild to moderate unilateral hearing losses.[1] The data analysis protocol calls for each psychophysical function in the experiment to be fit by an appropriate polynomial curve. Then, using the first derivative of the curve's equation, the slope of each curve in the data set is to be determined at several specific points along the curve. Collectively, these slope values provide the primary numerical data set for a variety of further statistical procedures concerned with characterizing the growth of loudness in hearing-impaired ears.

To obtain the necessary slope measurements, CD proceeded in a straightforward manner to use CA-Cricket Graph III 1.5 (1993), her trusty graphics application program for the Macintosh, to generate each curve in the data set and to obtain its equation. The third-order equation shown in Figure 1 is an example of the equations that CA-Cricket Graph III 1.5

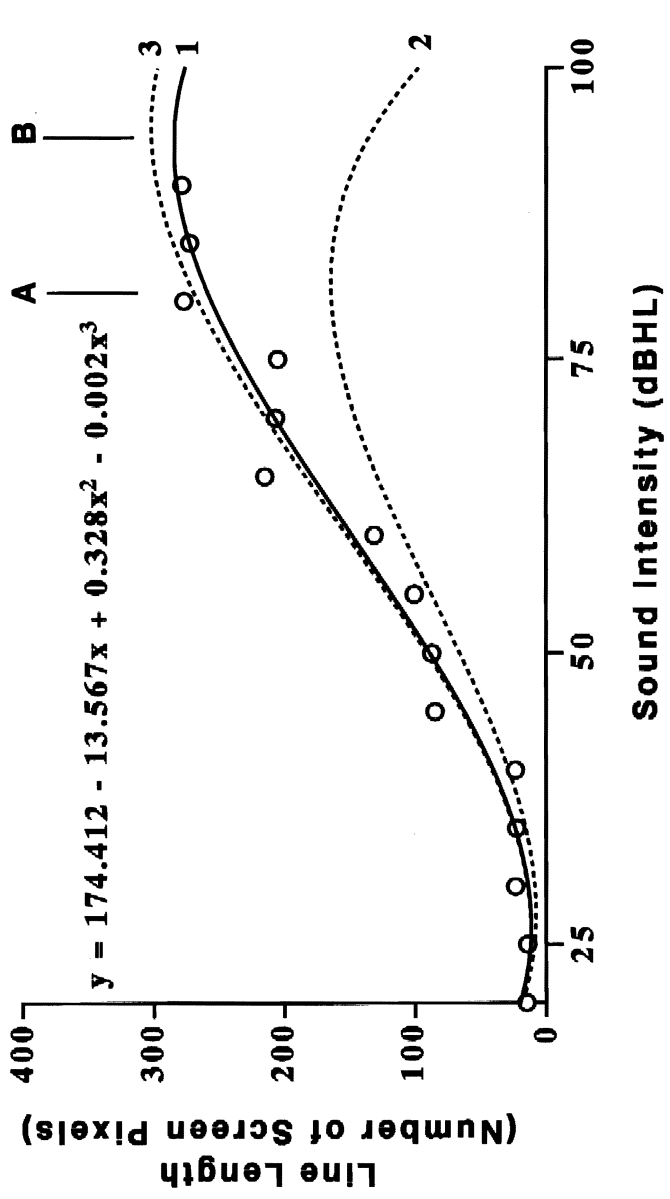


FIGURE 1 Plot produced by Cricket Graph III 1.5 of loudness against visual length scaling from a cross-modality matching study. Curve 1: Supplied by Cricket Graph based on 14-digit-precision polynomial coefficients; Curve 2: Supplied by the authors based on the equation supplied by Cricket Graph's default setting of three-decimal-place rounding for displayed polynomial coefficients; Curve 3: Supplied by the authors based on four-decimal-place rounding for the cubic term coefficients and three-decimal-place rounding for the other terms' coefficients.

supplied at the same time that it drew the curves for particular fits, such as curve 1 in the figure. The equation has three polynomial terms (or powers), namely the linear (x), quadratic (x^2), and cubic (x^3) terms, plus a constant. Naturally, one would think that the curve described by the equation and the curve plotted in the figure would be one and the same curve.

In fact, this is not the case. The equation, taken as written, actually generates curve 2 in the figure, a curve that veers far to starboard of the least squares solution curve over most of its course, and a curve that is not displayed by CA-Cricket Graph III 1.5. Whence the discrepancy? This errant second curve is the curve that one gets after rounding to three decimal places the 14-decimal-place coefficients of the three power terms that the least squares algorithm actually produced. By default CA-Cricket Graph III 1.5 simultaneously displays a curve based on double-precision calculations but an equation with coefficients rounded to three decimal places.[2] Consequently, the powers we see are not the powers that be, not precisely. It is not surprising that round-off error should lead to deviations of curves from their double-precision forms. However, it is counterintuitive, without thinking about it too deeply, that the deviation would be so systematic and extreme. Behavioral scientists are used to thinking of numerical round-off error in the data reduction process as having global and independent effects on data sets. For example, rounding the height and weight measurements of 20 people to the nearest inch and pound, respectively, should add noise to the relationship between height and weight, but should not grossly distort the general *form* of the relationship.

Not so when the round-off error is located in coefficients of polynomials rather than in data values themselves. In that case, the round-off error can significantly alter the relative contributions of the different terms. Generally, terms with higher powers are more affected by round-off error than terms with lower powers, and therefore terms with higher powers contribute more to the total measurement error than terms with lower powers. This error becomes obvious for numerically large abscissa values. For example, at a sound intensity of 85 dB HL, and with the CA-Cricket Graph III 1.5 default criterion of three-decimal round-off, the cubic component of the equation in Figure 1 is off by 9.66% of its true value ($0.002 \times 85^3 = 1228.250$ for three-digit round-off versus $0.0018... \times 85^3 = 1120.01$ for the maximum 14-digit precision of the coefficient), whereas the quadratic component is off by a mere .085% of its true value ($0.328 \times 85^2 = 2369.800$ for three-digit round-off versus $0.328..... \times 85^2 = 2371.825$ for the maximum 14-digit precision of the coefficient). The round-off error

associated with the cubic term, therefore, accounts for greater than 99% of the measurement error. Therefore, primarily because of the cubic term, the growth of absolute error as a function of increasing abscissa value is both substantial and orderly.

Although there is nothing mysterious in this realization once you think about the way the equations for polynomials work, it is not your foremost thought when you view that lovely tight-fitting curve alongside that tidy equation; and, in any case, three decimal places (that's thousandths!) certainly seems like quite a cautious round-off compromise. Ordinarily, you might never know that there was so much error in your calculations unless you were looking for it.

So, how did we come to appreciate the enormity of the error in this measurement protocol? It took a body slam to clue us in. After identifying a relevant point on one plot where the slope was obviously positive, CD calculated the slope's precise value (she thought) and obtained a negative number! Something was wrong. She brought this to the attention of the other one of us (VS). Together we concluded further that something was certainly wrong. And, within a half hour or so we had suspended our blind faith in the software long enough to discover what it was, namely that the powers that be, the higher they be, are relatively intolerant of error, even apparently small error, in their coefficients. This intolerance can lead to anomalous outcomes. Notice in Figure 1, for example, that in the region of the curves between the lines labeled A and B, the slope of the 14-digit-precision curve at any point is positive but the corresponding slope of the three-digit-precision curve is negative.

Having discovered the source of this error, CD probably would have just gotten on with it with new precision and vigor. However, it was impossible for us to avoid buttonholing a few colleagues that day to complain about those insidious polynomial coefficients. We polled people from a range of disciplines, including audiology, speech science, linguistics, psychology, engineering, physics, and mathematics (both applied and theoretical). To each we asked the following questions and received from each essentially the following answers:

1. Question: "Here is a plot of psychophysical data and the curve that our program fit to it, along with the equation that it displayed for the curve. Would you assume that the equation you see describes the curve you see?" Answer: "Sure. Doesn't it?"

2. Question: "Of course, the coefficients have been rounded off; from 14 to 3 digits. However, the curve is based on calculations that use the full

14 digits. Do you still think that the equation you see matches the curve reasonably well?"Answer: "Sure. Doesn't it?"

3. Question: "Do you think that, even given round-off error, if you were to plot the curve, using only the three-decimal-place coefficients, that it would tend to more or less overlap with the curve derived using 14-digit precision over the entire range from 20 to 100 dB HL?" Answer: "Sure. Wouldn't it?"

4. Question: "What if we told you that the second curve in the figure is the curve you actually get when you plot that equation as it is written with three-digit precision on the coefficient?" Answer: "No! Is it?"

5. Question: "Do you know why? The error on the cubic term has a huge effect because x^3 can get to be a huge number as x gets larger, so even a small coefficient error has a large, systematic effect over the course of the curve. Do you think you would have anticipated this if you were doing the curve fitting using a routine graphics package?" Answer: "Uh, ... yeah. Sure. See you later."

In fact, all of our casual informants believed that whatever error might be introduced by the rounding process would have the effect of causing the three-digit-precision curve to be only somewhat off from the 14-digit-precision curve. All claimed that the effect we were showing them was counterintuitive and that they would most likely have trusted the program to display an equation that would more accurately generate the curve it accompanied. Furthermore, they all initially expected the error to be unsystematically distributed over the entire course of the function, just as if it were random sampling error that were operating to cause the discrepancies. And, although our more mathematically oriented colleagues were quick to yawn out a perfunctory acknowledgment of the significance of round-off error in polynomial coefficients once we pointed it out to them, our colleagues in the behavioral sciences tended to have more of a "There but for the grace of God ..." attitude when they became aware of the surprise that we had experienced earlier due to our own blind faith. And, some shuddered with nervous recollection.

Rounding off polynomial coefficients to arbitrary numbers of decimal places, then, can lead to serious misrepresentation of the underlying form of the relationship between variables observed in psychophysical experiments. Specifically, the choice of an arbitrary round-off criterion for polynomial coefficients can produce ill-conditioned solutions, that is, solutions in which small changes due to error in the values of operands can lead to large changes in the calculated solutions of numerical algorithms

(Ortega, 1972, p. 3). Figure 1 clearly illustrates how round-off error can create an ill-conditioned solution to the curve-fitting problem. For example, the 14-digit-precision curve in Figure 1 (curve 1) is very closely approximated by a function that retains only four decimal places in the coefficient for the cubic term (curve 3), at least within the domain of the function populated with actual data observations. Note that curve 3 was drawn with the coefficient .0018 for the cubic term, whereas curve 2 was drawn with the coefficient .002 for the cubic term. That is, the rounding process incremented the coefficient only slightly, yet this small difference yielded the dramatic deviation illustrated by curve 2.

Ill-conditioned solutions due to the choice of arbitrary round-off criteria for polynomial coefficients can lead to severe mismeasurement and prediction error in experimental and clinical psychophysical activities. It is not a simple matter, however, to specify exactly how many digits it is necessary to retain in order to keep a polynomial solution well-conditioned. It is well known that many sources of error may interact to influence the accuracy of a numerical solution to the curve-fitting problem, and generally, these will interact in complex ways to cancel and amplify errors as they propagate through the many steps of a numerical algorithm or analysis protocol (Conte & de Boor, 1972, p. 10; Dahlquist & Bjorck, 1974, pp. 22-23). The exact number of digits needed for the coefficients of a given polynomial term will depend upon the specific function that characterizes the data, the order of the power for which the coefficient must be estimated, the abscissa domain over which it is desirable to maintain precision, the inherent precision with which data are represented by the computer or calculator that is being used, and so on.

The obvious solution to this round-off problem is not to round off. However, if it is desirable to round off and still maintain reasonable precision, then there are a couple of points to keep in mind. First, it is often said that the results of a numerical operation will not be any more accurate than the accuracy of the original data. However, this does not mean that coefficients or intermediate computational results can be rounded off to the number of significant figures in the data. Rather, most texts on numerical analysis advise that more significant figures be used throughout the computational process (Nielsen, 1967). The example presented in this paper shows that it is particularly critical that the rounding decision for polynomial coefficients not be based on the number of significant figures in the data set, especially for higher powers.

Second, the number of decimal places retained for polynomial coefficients should be commensurate with the powers of their polynomial terms. Generally, as the power of the polynomial term increases, more decimal places will be needed to maintain precision over the domain of the data set upon which the fitted curve is based.

Third, because of the complexities involved in specifying the exact effect of the different sources of error on the curve-fit process, careful attention should be given to establishing a protocol of checks on the accuracy of the results (Dahlquist & Bjorck, 1974, p. 23). We recommend that alternative round-off settings for specific equations be checked against the highest-precision values available on the computer or calculator by the simple exercise of plotting their curves all together on the same graph.

Heeding these simple rules of thumb will provide what computer cognoscente might refer to as the "WYSIWYG" precision guarantee for curve-fit displays, that is, the guarantee that "what you see is what you get!"

FOOTNOTES

[1] The long range goal of this work is to provide a valid response task to eventually study the growth of perceived loudness in a congenitally deaf population with severe to profound bilateral hearing losses. The initial study with individuals having unilateral hearing loss was designed to validate the cross modality matching technique by using individuals' normal ears as their own controls for performance in their impaired ears. Subjects were to draw a line on a computer CRT screen whose length indicated the loudness of a sound presented at a given intensity.

[2] This is an out-of-the-box default setting and is not a serious problem with CA-Cricket Graph III 1.5, which we find to be an excellent product. CA-Cricket Graph III 1.5 allows users to set much greater precision simply by selecting the equation in the graph window, opening its numerical format dialog box, and specifying the appropriate display format. Other graphing and statistical packages may have different out-of-the-box default settings. For example, JMP 3.0.2 (1994) has a default of nine decimal places, and can also be set to accommodate high round-off precision for particular polynomial curves.

ACKNOWLEDGEMENTS

This work was conducted at the National Technical Institute for the Deaf, a college of the Rochester Institute of Technology, in the course of an agreement with the U.S. Department of Education. We are grateful to all of our busy colleagues for graciously participating in our informal poll on round-off error in polynomial coefficients, and we respect their unanimous request to remain anonymous.

REFERENCES

CA-Cricket Graph III for Macintosh (Version 1.5) [Computer software]. (1993). Islandia, NY: Computer Associates International, Inc.

Conte, S. D., & de Boor, C. (1972). *Elementary numerical analysis* (2nd ed.). New York: McGraw-Hill Book Co.

Dahlquist, G., & Bjorck, A. (1974). *Numerical methods.* New Jersey: Prentice-Hall, Inc.

JMP Statistical Visualization Software (Version 3.0.2) [Computer software]. (1994). Cary, NC: SAS Institute, Inc.

Nielsen, K. L. (1967). *Methods in numerical analysis* (2nd ed.). New York: The Macmillan Co.

Ortega, J. M. (1972). *Numerical analysis: A second course.* New York: Academic Press.

JOURNAL OF OUTCOME MEASUREMENT®, 2(2), 168 -170

Book Review

Trevor Bond
James Cook University

McNamara, Tim. (1996) *Measuring Second Language Performance.* London: Longman. (Review copy provided by Addison Wesley Longman Australia Pty Ltd Price AUD 46.95 $US 35.70)

I think I'll scream if another colleague approaches me at a conference and enquires, "Can you tell me in five minutes what this Rasch analysis is all about?" For those of you who present to predominantly measurement audiences, the probability of being asked that question with the sincere expectation of a definitive answer by someone little versed in any of the minutiae of psychometrics is very slim indeed. But for those of you for whom the substantive area of research interest is something like school performance, intellectual development, rehabilitation, client satisfaction, language acquisition or the like and Rasch modelling is the tool by which the mysteries of these areas are laid bare, the prospect of repeatedly initiating small groups of students, institutional colleagues, clients and conference delegates to the strengths of Rasch modelling remains rather daunting. It is in that general context, as well in the specifically addressed content - the measurement of second language performance - that Tim McNamara's book should have substantial beneficial impact.

The plan of McNamara's work revolves around three interrelated and often revisited components: a conceptual analysis of what constitutes *second-language performance;* how that performance can be *measured,* and how the various members of the Rasch family of models are appropriate to his self imposed task. While Rasch modelling is widely accepted in Australia as the preferred analytical method in language assessment (and many educational settings), the author is careful to outline the distinctive features of language performance and patiently to address how the Rasch models lend themselves to the meaningful solution of the difficulties encountered by practicioners. The culmination of this development of key ideas focuses on the development of the *Occupational English Test* in Chapter Four.

Interestingly, McNamara was introduced to Rasch analysis by Geoff Masters whose closing remarks at the 1987 AILA World Congress in Sydney predicted, "As language testers become increasingly familiar with these methods of analysis *(IRT and the Rasch models in particular)*, we can expect to see these methods contributing to an improved understanding of the nature of developing language proficiencies as well as being used to study and understand the language development of individual learners." It seems that at once McNamara's book is the culmination of the promise expressed in Masters's paper as well as a device by which that promise might be expanded and fulfilled in the area of assessing language proficiency.

While the book has very distinctive claims to be an original and considerable contribution to the assessment of language proficiency, my personal interest in the text is two-fold. The first is the particular interest the book should be to the general Rasch measurement fraternity in terms of the very clear and practical explanation of fundamental Rasch concepts which comprises the second half of the book. While the focus is clearly on language assessment, the presentation of the 'new measurement' concepts assumes that the reader has little prior knowledge of theories of measurement. As a trial of its utility in this regard I gave the text to a MEd student interested in the Rasch modelling of longitudinal cognitive developmental data she has collected of her secondary school students over a five year period. The clarity of her understanding of Rasch principles was due in no small part, I believe, to the detailed discussion of them by McNamara. In this sense, it is an indispensable addition to the library of any Rasch-oriented university teacher.

Secondly, McNamara's account remains very clear on the distinction between the complex theoretical underpinnings of second language acquisition that guide his investigations and the meaning that may be attributed to the Rasch modelled empirical data gleaned from performance testing. In an empiricist-dominated measurement culture that seems to take the *wysiwyg* precept as its touchstone, even amongst apparently well-read Rasch theorists, such carefully drawn and illustrated distinctions as that between 'psychological unidimensionality' and the 'psychometric unidimensionality' fundamental to Rasch modelling is both refreshing and informative. The author warns us (p.269), "The point is that interpretation of the results of Rasch analysis must be informed by an in-principle understanding of the relevant constructs." The discussion of unidimensionality that follows is strongly recommended as required reading by both Rasch

modelling adherents and their critics.

So while we might have to turn elsewhere for something akin to the 10 minute shorthand summary for novices of why Rasch modelling works so well (e.g., Smith, et al, 1997), we can confidently rely on McNamara for a well couched initiation to the detail of Rasch modelling for the newcomer who really wishes to try the techniques. Moreover, those interested in well-informed accounts of empirical aspects of second language performance should regard this book as indispensable.

REFERENCES

de Jong, J. H.A. L. & Stevenson, D. K. (Eds.). (1990). *Individualizing the assessment of language abilities.* Clevedon: Multilingual Matters.

Smith, L., Dockrell, J. & Tomlinson, P. (Eds.). (1997). Chapter 10 Measuring development. *Piaget, Vygotsky & beyond.* London: Routledge.

CONTRIBUTOR INFORMATION

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

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

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

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

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

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

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

SUBSCRIBER INFORMATION

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

 ISSN 1090-655X.