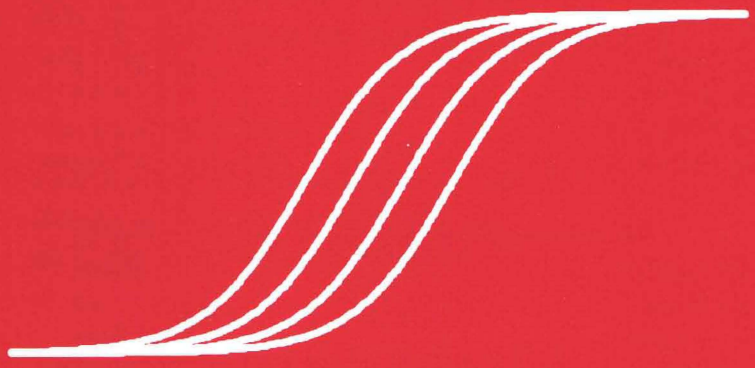
Volume 1, Number 3, 1997 ISSN 1090-655X

Journal of

Outcome Measurement

Dedicated to Health, Education, and Social Science



REHABILITATION FOUNDATION INC.

Est. 1993

Research & Education

JOURNAL OF OUTCOME MEASUREMENT

Volume 1, Number 3 1997

JOURNAL OF OUTCOME MEASUREMENT, *1*(3), 173 -191

On-line Performance Assessment Using Rating Scales

John Stahl
and
Rebecca Shumway
Computer Adaptive Technologies

Betty Bergstrom
American Dietetic Association

Anne Fisher
Colorado State University

The purpose of this paper is to report on the development of the on-line performance assessment instrument - the Assessment of Motor and Process Skills (AMPS). Issues that will be addressed in the paper include: (a) the establishment of the scoring rubric and its implementation in an extended Rasch model, (b) training of raters, (c) validation of the scoring rubric and procedures for monitoring the internal consistency of raters, and (d) technological implementation of the assessment instrument in a computerized program.

Requests for reprints should be sent to John Stahl, Computer Adaptive Technologies, Inc., 2609 West Lunt Avenue, Suite 2E, Chicago, IL 60645.

INTRODUCTION

The use of performance assessment has increased dramatically in the past few years in many different arenas of testing. This growing interest in the benefits of performance assessment has developed concurrently with widespread criticism of more objective assessment tools, particularly multiple choice tests. Some of the most frequently addressed flaws in objective testing are that they utilize a one-right-answer approach, draw on a narrowed curriculum, and primarily address only discrete skills (Hambleton & Murphy, 1992). Educators are turning to performance assessment as a means of achieving more "authentic" measures of a student's ability to perform a task or mastery of a subject; job performance evaluations are relying more on observational data in assessing the competence of employees, and certification and licensure organizations are integrating task related performance assessments, sometimes referred to as work samples, into their batteries of test instruments.

The time and labor costs involved in implementing a performance assessment, however, are major considerations that must be taken into account (Reckase, 1993). In a typical performance assessment, examinees prepare work samples which are collected at a central location. Raters or judges are assembled, undergo a training session that can vary in intensity, and then rate the quality of the samples. Alternatively, judges are trained and sent out to observe and assess the performance of examinees. In addition, the development and validation of a scoring rubric involves an enormous investment of time and effort, including input of expert judgment, field testing of preliminary drafts, and finalization of the evaluation instrument.

On-line performance assessment was developed to maximize the utility of performance assessment and to minimize the time and labor costs incurred. A computerized assessment software program was designed to provide a convenient tool for raters entering observational rating scores and to produce meaningful feedback on clients' progress and the quality and consistency of raters. Using previously collected data on items to be assessed, difficulty calibrations of these items are established in accordance with an extended version of the Rasch model. These calibrations are fundamental to the on-line assessment. They provide calibrated standardized cases for the training of raters and establishment of rater severity calibrations. The input of these item difficulty calibrations and rater severity calibrations into the on-line assessment program make it

possible to estimate client ability.

A series of reports provides information regarding client performance, areas in which improvement is needed, and unexpectedly high or low scores--which may be indicative of errors in score recording or rater inconsistency. The computerized system can also produce a report displaying the results of multiple evaluations, allowing the rater to analyze progress over time. Together, these reports enable raters to direct their efforts toward areas in which clients have the greatest need. In addition, because the on-line system utilizes full password protection, confidentiality of the information it contains is guaranteed.

The purpose of this paper is to report on the development of the on-line performance assessment instrument - the Assessment of Motor and Process Skills (AMPS) (Fisher, 1995). Issues that will be addressed in the paper include: (a) the establishment of the scoring rubric and its implementation in an extended Rasch model, (b) training of raters, (c) validation of the scoring rubric and procedures for monitoring the internal consistency of raters, and (d) technological implementation of the assessment instrument in a computerized program.

THE AMPS PROJECT

AMPS is a performance assessment instrument used by occupational therapists to determine the effectiveness of a program of rehabilitative therapy for clients. At various stages in the therapeutic regime, the therapist observes the client performing domestic or instrumental activities of daily living (IADL) and evaluates the client's performance. The goal is to determine whether or not the intervention is assisting the individual in his or her ability to function independently, efficiently and safely. Clients are rated while performing standardized instrumental activities of daily living (IADL), such as making a peanut butter and jelly sandwich or sweeping the floor. Two categories of rating scales are used: motor skills and process skills.

The AMPS rating scales were developed by Anne Fisher at Colorado State University (Fisher, 1995). The administration of an AMPS evaluation encompasses the following steps:

1. The client is interviewed and then chooses two or three of the AMPS IADL tasks to perform.

2. The client and the therapist agree on any constraints of the tasks chosen. This includes such things as the specific ingredients the person plans to use.

3. The client and the therapist set up the environment in which the tasks are to be performed.

4. The client and the therapist review the task contract. The AMPS observation begin.

5. The therapist observes the client's performance of the tasks.

6. The therapist scores the performance. For each task performed, the client is evaluated on 16 motor skill items and 20 process skill items. The therapist has the option of directly entering the scores into the AMPS computer scoring program (Fisher, 1995, p.3).

Motor skills are those observable actions that a person uses to, "move oneself or the task object during all task performances," while process skills are those actions that a person uses to, "sensibly organize and adapt the actions of task performance as the process unfolds over time" (Fisher, 1995, p.3). A listing of the motor and process skills are found in Table 1. A detailed discussion of each of these skill areas and how they are used in an AMPS evaluation can be found in Fisher (1995).

The ultimate goal of the AMPS performance assessment is to answer two questions: (a) "Why does this person experience difficulty?" and (b) "What level of task challenge can this person handle?" (Fisher, 1995, p. 4). The first question is answered by examining the ratings assigned to the various tasks and identifying those particular skill areas or tasks that are especially difficult for the individual to perform. This examination provides the groundwork for planning future interventions to address specific areas of disability experienced by the client. The second question is answered by analyzing the ratings using the extended Rasch model, or FACETS model (Linacre, 1988; Linacre 1989). This analysis generates client ability measures located on two scales of ability: one for motor ability and one for process ability. The scales are linear, allowing the therapist to gauge the client's progress toward recovery. By making repeated observations during the course of therapy, the therapist is able to determine the effectiveness of the interventions.

Table 1
Motor and Process Skills Used in AMPS

Motor Skills		
Stabilizes	Coordinates	Lifts
Aligns	Manipulates	Calibrates
Positions	Flows	Grips
Walks	Moves	Endures
Reaches	Transports	Paces
Beads		
Process Skills		
Paces	Initiates	Restores
Attends	Continues	Navigates
Chooses	Sequences	Notices\Responds
Uses	Terminates	Accomocates
Handles	Searches/Locates	Adjusts
Heeds	Gathers	Benefits
Inquires	Organizes	

Scoring Rubric

The AMPS scoring rubric uses a 4-point rating scale to codify the observations of the performed skills on the agreed upon tasks. A score of "4" indicates effective performance of the skill, whereas a score of "1" indicates extreme difficulty (unsafe practice or need for assistance) in performing the task under observation. The results of the scoring are analyzed using an extended version of the Rasch model. The basic Rasch model can be considered a two-facet model with one facet for person ability and a second facet for item difficulty, whereas the extended Rasch model is a latent trait model which estimates multiple facets--in this case, client ability, rater severity, task difficulty, and skill item difficulty. The equation for the extended Rasch model used in the AMPS program is as follows:

$$\mathrm{Ln}\left(\frac{P_{ntisk}}{P_{ntis(k-1)}}\right) = B_n - D_t - E_i - R_s - F_k \qquad (1)$$

Where: P_{ntisk} = the probability of awarding a value on the rating scale step k.

$P_{ntis(k-1)}$ = the probability of awarding a value on the rating scale step k-1.

B_n= the ability of the client to perform the observed skill as a part of the agreed upon task.

D_t = the difficulty of the task being observed.

E_i = the difficulty of the skill item being observed.

R_s = the severity of the rater performing the observations.

F_k = the difficulty of that particular score or step on the rating scale.

Each component of the assessment is modeled as an independent facet. Facets for client ability, rater severity, task difficulty, item difficulty, and scale step difficulty are constructed. For each facet, the sum of the scores awarded is used to estimate the placement of facet elements on a common scale.

A standard error of measure accompanies each facet, indicating the precision of the estimate. As with all measurements, an increase in the number of observations increases the precision of the estimation. All estimates are on a common log-linear scale, which allows comparisons to be readily made. Interactions can also be modeled, allowing detection of unusual interactions between raters and skill items, or raters and particular clients.

In addition, two fit statistics are routinely presented with each analysis, identifying any client, rater, task, or item whose participation in the rating assessment deviates from the expectations of the model. In general terms, the expectation is that the more able clients will score higher on the tasks and items than less able clients; that more difficult tasks and items will receive lower scores than easier tasks and items; and that more severe raters will award lower scores than more lenient raters.

The infit statistic is the weighted mean-squared residual across all cases, weighted by the variance of the probability of achieving a certain score. It is sensitive to deviations at the points close to the center of the scale. The outfit statistic is the mean-squared residual across all cases. Because it lacks the weighting of the infit, it is more sensitive to outliers which appear as unexpectedly high or low ratings. As such, the outfit

statistic provides a particularly valuable clue in detecting errors in the application of the rating scale. It is also valuable in highlighting especially strong or weak areas in a client's therapeutic program.

The Training of AMPS Raters

The usual approach to evaluations that use raters is to view any variability between ratings as an undesirable source of error. The goal of almost all training programs is to reduce rater variability as much as possible. For example, if the "validation committee" determines that a simulated client should receive a score of "3" on a particular skill on a designated task, then all raters are "trained" to assign a "3" to that observation. In addition, scoring rubrics are refined to reduce instances where raters might tend to disagree. In a rather procrustean manner, rater training and rigidly structured scoring rubrics are used in an attempt to force raters to emulate an "ideal" rater. The rather poor inter-rater reliability statistics generally reported attest to the lack of success of such procedures (Stahl and Lunz, 1996).

By contrast, the training of AMPS raters takes a more Pollonian approach, in that raters are trained to be as honest and consistent as possible in their assessment of their clients. In other words, in the AMPS training program, variability between raters is accepted as a given. This is not to say that raters are allowed complete freedom; they receive extensive training in the concepts underlying each of the skills being rated and are presented with opportunities to apply these concepts in actual evaluations. The training of an AMPS rater extends over 5 days. During this training, the following points are covered: the establishment of the task contract with the client, the meaning of each of the skills observed in the context of the client's overall performance, the use of the 4-point rating scale in establishing the observational score, the use of the AMPS computer program, and interpretation of the computerized output. The therapists become very familiar with the evaluation rubric, but no attempt is made to force them to use the rubric in the same manner as any other therapist. As discussed above, rater variability is accepted as a latent trait in the extended Rasch model. Likewise, variation in rater severity is modeled as part of the estimation equation. Therefore, the resulting measures for a client take into account these differences among therapists.

During the course of their training, therapists are asked to rate a variety of pre-calibrated cases. These cases are presented on video tape and

function as standardized cases. The estimated ability of the clients represented in these cases has been established by collecting observations over an extended period of time and drawing on a large number of therapists. As of 1995, more than 5,000 clients from North America, Scandinavia, the United Kingdom and Australia had been used in the AMPS development. A total of 56 tasks and 36 skill items have been calibrated and found to fit the extended Rasch model (See Equation 1). Close to 500 therapists were used in the calibration studies (Fisher, 1995, p. 123).

Once a therapist has completed the training course, the ratings collected for him or her during the course of the training are used to generate a preliminary rater severity calibration. This rater severity is based on Equation 1. All elements on the right side the equation, except the rater severity, are fixed values. The probabilities for the left side of the equation are derived from the ratings that the therapist awards the standardized cases. Once the rater severity calibration has been constructed, the therapist is required to collect data on an additional 10 clients in his or her normal therapeutic setting. The results of these observations are then forwarded to the AMPS Project. There, a refined rater severity will be generated and an update on the rater's severity will be sent to the rater to be incorporated in his or her copy of the AMPS scoring program.

Validation of the AMPS Program

Prior to the implementation of the AMPS program in an on-line computer product, an extensive validation program was undertaken. The analysis of the AMPS rubric for the purpose of developing computer-scoring software was accomplished in two stages. First, existing data was analyzed to, "verify that a single, international, cross-cultural scale could be developed and used to assess clients from diverse diagnostic subgroups" (Fisher, 1995, p. 127). Approximately 3,000 clients were used in the initial study. The primary purpose of the study was to verify the stability of the task and skill item difficulties across different subgroups, based on either gender, ethnic or diagnostic categories. Inconsistencies in difficulty estimates across subgroups can affect client ability estimates produced by the computerized program. Some variability was detected in the item difficulties across diagnostic groups; however, when client ability estimates were derived using the subgroup specific item difficulties as

opposed to a total group item difficulties, the researchers found that, "approximately 95% of the subject ability measures remained stable within ±0.10 logits and that less than 1% differed by more than the mean standard error" (Fisher, 1995, p. 130). (The reader is referred to Fisher, 1995 for a complete description of the validation studies.) In the final step of the validation process, data from 4,766 clients were used in a final calibration run. From this analysis, calibration values were derived for each of the 16 motor skill items, 20 process skill items, and 56 tasks that were used in the development of the AMPS computer scoring program.

IMPLEMENTATION OF THE AMPS PROGRAM ON COMPUTER

The AMPS computer scoring program was developed as a joint effort of the AMPS Project and Computer Adaptive Technologies, Inc. From the beginning, several principles guided the program's development. The scoring program had to be relatively simple to use, particularly for the input of the observational rating scores. Also, the program had to produce output that would contribute meaningfully to the progress of the therapy while providing the therapists feedback that would allow them to monitor their own internal consistency and highlight areas of concern.

The foundation of the program was the data previously collected by the AMPS Project. Based on this data, calibrations on the tasks and skill items to be used in the software program were established. These calibrations also provided the standardized cases used in the training of therapists as AMPS raters and were entered into a separate AMPS rater severity calibration program as fixed calibrations to be used in the estimation of rater severity. The construction of client ability measure paralleled the way that raters were calibrated, except that in this case all values on the right side of the basic equation are fixed except for the client ability (See equation above). Again, the probabilities are derived from the rating awarded and the degree of rater severity is accounted for in the estimation process.

The AMPS computer scoring program is pre-loaded with the severity of the raters and the difficulty of the tasks, skill items, and rating steps. Demographic information on the client, such as age, sex, ethnic group, and medical diagnosis, also can be pre-loaded. With this information contained in the program, a therapist merely has to indicate which task or tasks are being performed and enter the ratings awarded to the various skill items when performing an evaluation. The program is setup to allow the entering

the ratings, using a data entry screen, immediately after the observations are conducted. Also, a provision is made so that free-form notes can be attached to any evaluation, allowing the therapist to include observations that fall outside the structure of the scoring rubric, but may be highly pertinent to the conduct of the therapy.

Reports

A set of five preestablished reports are integral to the program. (Examples of each report can be found in Appendix A.) The battery of reports generated by the AMPS program provides the therapist with the information necessary to guide the course of a client's intervention and to self-monitor his or her own internal consistency. The AMPS Report categorizes the ratings for each skill item into a three-step ordinal scale: Adequate, Marginal, and Markedly Deficient. This report provides a summary of the client's performance and readily highlights areas in which the client demonstrated a need for further improvement.

The Misfit Report is designed to highlight unexpectedly high or low scores awarded to a skill item or items. Unexpected high or low scores may result from errors in data entry of scores. More importantly, they also can be the result of either a client deficiency that requires special attention, or a rater's internal inconsistency. In the case of a client deficiency, the Misfit Report can help the therapist to focus his or her efforts where they are most required. In the case of rater inconsistency, the report assists the therapist in monitoring his or her use of the rating scale. If a rater consistently misfits on a particular item, for example, then he or she can begin to explore the reasons why he or she has a tendency to award either higher or lower scores on that particular item. The Misfit Report is an advantage of the extended Rasch analysis in that it provides much more detailed diagnostic information than other methods of analysis that focus on aggregate scores. In addition, the report presents information in a non-comparative, non-judgmental manner, which has been found to be more favorable to raters than reports which compare their performance to other raters (Stahl and Lunz, 1996).

The Notes Report allows the therapist to print out a copy of any notes recorded as part of an evaluation. Notes serve as a valuable adjunct to the rating evaluation and allow the therapist to record observations in a free-form memo format. An optional title can be attached to the notes report to enable the therapist to keep track of the reports printed.

The fourth report is the Graphic Report which displays the estimated ability measures of the client on both the motor and process scales. The greatest benefit of the Graphic Report is that the results of multiple evaluations can be displayed on the same report. If the evaluations shown on the report were conducted at different times, then the amount of progress made between the two evaluations is readily apparent on the report. This allows the therapist to easily gauge the measured effectiveness of the interventions conducted between the two occasions. An added feature of the Graphic Report is its indication of a cut-off measure on both scales. (Clients who are below the cut-off in ability measure on the motor scale are judged to have motor deficits affecting the degree of effort exerted. Clients who are below the cut-off ability measure on the process scale are judged to require assistance for community living.) These cut-off scores are based on thousands of empirical observations of clients' functional ability.

The final report, the Raw Scores Report, is a print out of the raw scores awarded to each skill item on each task observed during a particular evaluation. The Raw Scores Report provides a hard copy of the evaluation observations which can be incorporated into a client's file and used as a record and reference for his or her course of therapy.

Confidentiality

Another important aspect of the software program is its password protection capability, which affords the highest confidentiality to information stored on clients. Only therapists who have been through the AMPS training program receive the rater disk that allows access into the AMPS program. Each therapist is assigned a password code that must be entered each time he or she accesses the program. Confidentiality of client data also is maintained when data is exported to the AMPS headquarters. All identification on the client is removed from the export files when they are created.

DISCUSSION

This paper has focused on the development of an on-line computerized assessment software program which implements the Assessment of Motor and Process Skills (AMPS) evaluation system. The AMPS program encapsulates the results of years of research into a highly portable, user-

friendly software system to provide therapists with the information they need to manage therapeutic regimes for their clients. While the AMPS program is specialized to meet the needs of the AMPS project, the steps taken in the program's development are directly relevant to the development of a similar on-line system for any program that uses raters for assessment purposes.

The developmental stages of performance assessment, such as the creation and validation of the scoring rubric and the training of raters, are vital components of both traditional and computerized assessment methods. But because the computerized system incorporates the extended Rasch model, it allows administrators of performance assessments to maximize the results of their assessments and to implement them with relative ease. By allowing each facet of the evaluation process to be independently estimated, the extended Rasch model permits the focus of the assessment to be centered appropriately, while all other facets of the evaluation process are fixed. The significance of this capability is two-fold: it allows the calibration of new raters during rater training, and it allows the estimation of client ability immediately following an evaluation.

The practical applications of on-line computerized assessment are broad. A prime example of an arena that could benefit from computerized assessment is the grading of essay tests. Using the on-line system, data could be evaluated using the extended Rasch model to yield calibrations on essay prompts and on types of essays. (See, for example, Engelhard, 1992 & 1994.) These calibrated essays, along with a set of calibrated samples, could then form the basis of rater training and the calibration of rater severity. In addition, a teacher, as a trained and calibrated rater, could take the computerized system into the classroom and perform essay evaluations that would provide immediate feedback to both the student, the school administrators, and the teacher. Other fields of assessment that could benefit from computerized assessment include competitions that use judges, performance evaluations in the work place, certification and licensure tests in which demonstration of ability to perform a given task is more important than test performance, and mastery tests in which educators must determine whether a student has mastered a subject.

REFERENCES

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

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

Fisher, A. G. (1995). *Assessment of Motor and Process Skills* (Rev. ed.). Fort Collins, CO:Three Star Press.

Hambleton, R. K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5*(1), 1-16.

Linacre, J. M. (1988). *FACETS: Computer program for many-faceted Rasch measurement* [Computer program]. Chicago:MESA Press.

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

Reckase, M. D. (1993, April). *A theoretical prediction of measurements properties.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Stahl, J.A., & Lunz, M.E. (1996) Judge Performance Reports: Media and Message, in *Objective Measurement; Theory into Practice, Vol. 3*, G. Engelhard & M. Wilson (Eds.). Norwood, NJ:Ablex Publishing Corp.

APPENDIX
Sample Reports

The following sample reports are included:

1) Assessment of Motor and Process Skills Report
2) Misfit Report
3) Notes Report
4) Graphic Report
5) AMPS Raw Scores Report

ASSESSMENT OF MOTOR AND PROCESS SKILLS

Client: MR. STAHL **Therapist:** ANNE FISHER
Id: 3334 **Gender:** Male
Age: 45 **Evaluation** 04/22/97

The Assessment of Motor and Process Skills (AMPS) was used to determine how MR. STAHL's MOTOR and ORGANIZATIONAL/ADAPTIVE (process) capabilities affect MR. STAHL's ability to perform functional DAILY LIVING TASKS necessary for COMMUNITY LIVING. The tasks were chosen from a list of standard functional activities rated according to their level of complexity. MR. STAHL chose to perform the following tasks which MR. STAHL considered to be meaningful and necessary for functional independence in the community:

Task 1: A-2 Hot or cold instant drink

The level of complexity of the tasks chosen was easier than average. Overall performance in each skill area is summarized below using the following scale: ADEQUATE SKILL: no apparent disruption was observed; DIFFICULTY: ineffective skill was observed; or MARKEDLY DEFICIENT SKILL: observed problems were severe enough to be unsafe or require therapist intervention.

The following strengths and problems were observed during the administration of the AMPS:

Adequate = A Difficulty = D Markedly Deficient = MD

MOTOR SKILLS:

Skills needed to move self and objects.

	A	D	MD
Posture:			
STABILIZING the body for balance.	X		
ALIGNING the body in a vertical position.	X		
POSITIONING the body or arms appropriate to the task			X
Mobility:			
WALKING: moving about the task environment (level surface)			X
REACHING for task objects.		X	
BENDING or rotating the body appropriate to the task		X	
Coordination:			
COORDINATING two body parts to securely stabilize task objects	X		
MANIPULATING task objects.			X
FLOWS: executing smooth and fluid arm and hand movements		X	
Strength and Effort:			
MOVES: pushing and pulling task objects on level surfaces or opening and closing doors or drawers.		X	
TRANSPORTING task objects from one place to another		X	
LIFTING objects used during the task.			X
CALIBRATES: regulating the force and extent of movements			X
GRIPS: maintaining a secure grasp on task objects		X	
Energy:			
ENDURING for the duration of the task performance	X		
Maintaining an even and appropriate PACE during task performance		X	

ASSESSMENT OF MOTOR AND PROCESS SKILLS

Client:	MR. STAHL	**Therapist:**	ANNE FISHER
Id:	3334	**Gender:**	Male
Age:	45	**Evaluation**	04/22/97

Adequate = A Difficulty = D Markedly Deficient = MD

PROCESS SKILLS:
Skills needed to organize and adapt actions to complete a task.

	A	D	MD
Energy:			
Maintaining an even and appropriate PACE during task performance		X	
Maintaining focused ATTENTION throughout the task performance.	X		
Using Knowledge:			
CHOOSING appropriate tools and materials needed for task performance.		X	
USING task objects according to their intended purposes.		X	
Knowing when and how to stabilize and support or HANDLE task objects.			X
HEEDING the goal of the specified task.			X
INQUIRES: asking for needed information.		X	
Temporal			
INITIATING actions or steps of task without hesitation.	X		
CONTINUING actions through to completion.		X	
Logically SEQUENCING the steps of the task.	X		
TERMINATING actions or steps at the appropriate time.		X	
Space and Objects:			
SEARCHING for AND LOCATING tools and materials.			X
GATHERING tools and materials into the task work space.	X		
ORGANIZING tools and materials in an orderly, logical, and spatially appropriate fashion.		X	
RESTORES: putting away tools and materials or straightening the work space.	X		
NAVIGATES: maneuvering the hand and body around obstacles.		X	
Adaptation:			
NOTICING AND RESPONDING appropriately to nonverbal task-related environmental cues.		X	
ACCOMMODATES: modifying ones actions to overcome problems.		X	
ADJUSTS: changing the work space to overcome problems.	X		
BENEFITS: preventing problems from reoccurring or persisting.			X

MISFIT REPORT

Client	MR. STAHL	**Therapist:**	ANNE FISHER
Id:	3334	**Gender:**	Male
Age:	45	**Evaluation Date:**	04/22/97

The following misfitting ratings were noted

The item score for Stabilizes was unexpectedly high on the task A-2 Hot or cold instant drink.

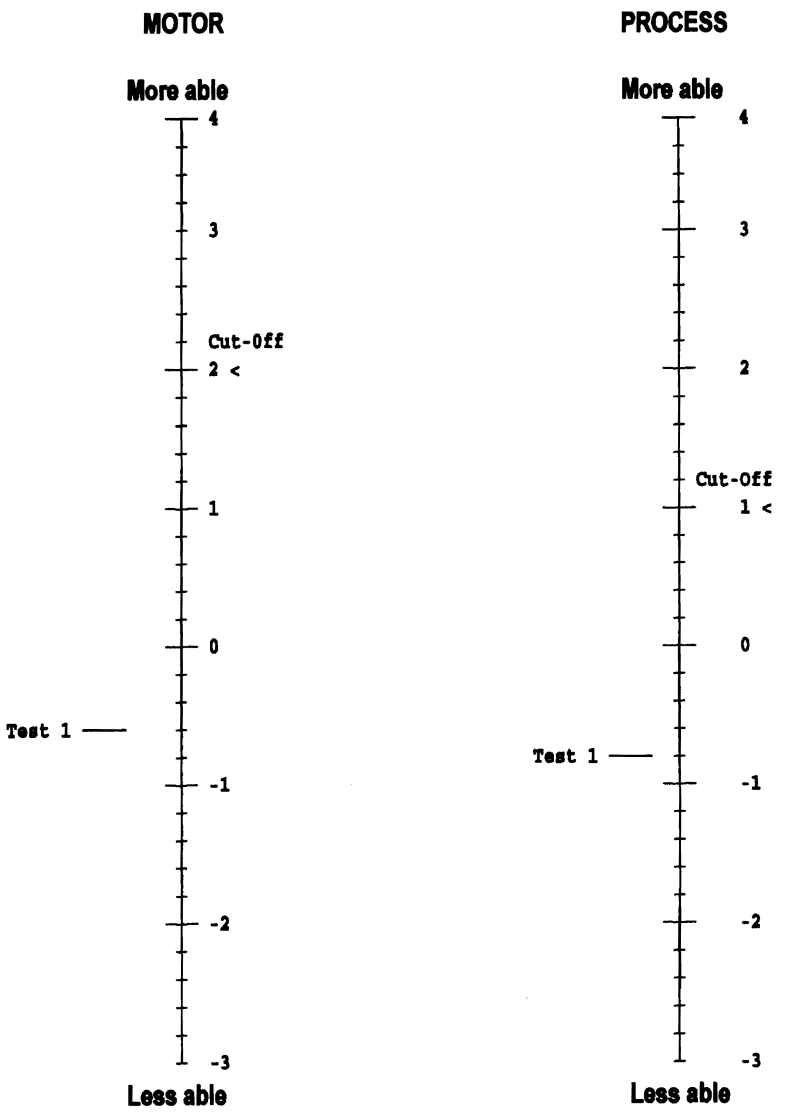
Refer to the AMPS manual for further information regarding possible reasons for the misfit.

SAMPLE NOTES

Client:	MR. STAHL	Therapist:	ANNE FISHER
Id	3334	Gender:	Male
Age:	45	Evaluation Date:	04/22/97

Client was very interested in the entire procedure and very cooperative. Showed unexpected ability on stabilizes.

GRAPHIC REPORT OF MR. STAHL'S AMPS RESULTS



MR. STAHL'S AMPS motor and process ability measures plotted inreference to AMPS scalecut-off measures indicative of evidence of problems that impact on performance.

	MOTOR	PROCESS
Test 1	-0.6	-0.8

AMPS RAW SCORES

Client: MR. STAHL
Id: 3334
Age: 45

Therapist: ANNE FISHER
Gender: Male
Evaluation Date: 04/22/97

Task 1: A-2 Hot or cold instant drink

MOTOR SKILLS

Posture:	Task 1
Stabilizes:	4
Aligns:	3
Positions:	1
Mobility:	Task 1
Walks:	2
Reaches:	2
Bends:	2
Coordination:	Task 1
Coordinates:	3
Manipulates:	2
Flows:	2
Strength and Effort:	Task 1
Moves:	3
Transports:	2
Lifts:	2
Calibrates:	1
Grips:	2
Energy:	Task 1
Endures:	3
Paces:	2

AMPS RAW SCORES

Client:	MR. STAHL	**Therapist:**	ANNE FISHER
Id:	3334	**Gender:**	Male
Age:	45	**Evaluation Date:**	04/22/97

PROCESS SKILLS

Energy:	Task 1
Paces:	2
Attends:	3
Using Knowledge:	Task 1
Chooses:	2
Uses:	2
Handles:	2
Heeds:	1
Inquires:	2
Temporal Organization	Task 1
Initiates:	3
Continues:	2
Sequences:	3
Terminates:	2
Space and Objects:	Task 1
Searches/Locate	1
Gathers:	3
Organizes:	2
Restores:	3
Navigates:	2
Adaptation:	Task 1
Notices/Respond	2
Accommodates:	2
Adjusts:	3
Benefits:	1

JOURNAL OF OUTCOME MEASUREMENT, *1*(3), 192- 218

Evaluating the FONE FIM: Part I. Construct Validity

Wei-Ching Chang
University of Alberta

Susan Slaughter and Deborah Cartwright
Northern Alberta Regional Geriatric Program, Glenrose Rehabilitation Hospital

Chetwyn Chan
Hong Kong Polytechnic University

Rasch analysis was used in this paper to evaluate the Motor component of the FONE FIM, the telephone version of the Functional Independence Measure (FIM). For this purpose, 132 patients discharged from an inpatient geriatric assessment and rehabilitation program were assessed by trained research assistants using the FONE FIM. The results at 5 weeks post-discharge were compared to the observation FIMs (OBS FIMs) done at home 6 weeks post-discharge. These patients had an average age of 79 years and presented with multiple, complex medical problems and significant functional decline. The FONE FIM and the OBS FIM were shown to share a strikingly similar item hierarchy, based on Rasch item difficulty measures. Only bladder management and climbing stairs were misfitting items as indicated by item fit statistics. The same 13-item set and 4-point scales were shown to be psychometrically optimal for both the FONE FIM and the OBS FIM based on the person separation index. Further research is required to address the issue of the optimal item set and scale levels from psychometric and clinical perspectives.

Requests for reprints should be sent to Wei-Ching Chang, Department of Public Health Sciences, Faculty of Medicine and Oral Health Sciences, 13-103 Clinical Sciences Building, University of Alberta, Edmonton, Alberta, Canada T6G 2G3.

An important goal of health care services for older adults is to facilitate more effective ADL function. Monitoring functions beyond the treatment period is essential for the measurement of functional outcomes. This can be accomplished by conducting post-discharge follow-up assessments.

Options for gathering follow-up data include face-to-face, telephone, or mail contact. Information may also be obtained from either the patient or a proxy respondent who is knowledgeable of the patient's condition. Each option has its strengths and weaknesses in terms of response rate, turnaround time, completeness of data, bias, burden, errors of interpretation, and cost (Guyatt, 1993; Smith, 1992; Weinberger et al., 1996). Not all options are appropriate for a specific assessment tool. Certain tools are designed only for trained raters, thus precluding self-administration through a mail survey. Such is the case with the Functional Independence Measure (FIM), which is used as an outcome measurement tool in the Northern Alberta Regional Geriatric Program in Edmonton, Alberta, Canada. The focus of our investigation was the extent to which the telephone mode may be used in lieu of the observational mode.

The FIM has been tested extensively in inpatient settings based on the observational mode, demonstrating a satisfactory level of face, content, and construct validity (Dodds et. al., 1993; Granger et al., 1990; Linacre et al., 1994), intra- and inter-rater reliability (Dodds et. al., 1993; Hamilton et al., 1994; Ottenbacher et al., 1996), sensitivity (Kidd et al., 1995) and feasibility (Kidd et al., 1995). The telephone version, the FONE FIM (Smith et al., 1990), has been in use since about 1990 by a number of institutions at a 3-6 month follow-up. This version, which was shown to agree fairly well with the observational FIM based on the raw scores (Smith et al., 1996), has also been shown to result in a high response rate with a reasonable cost (Smith, 1992). However, its validity has yet to be reported in the literature. Hence, an attempt was made to address this issue. For reasons that will become apparent, this paper will focus on the Motor component of the FONE FIM.

The following hypotheses were tested:

1. The 13 items in the Motor component of the FONE FIM constitute a single dimension.
2. The hierarchical structure of the FONE FIM is comparable to that of the OBS FIM (the observational or face-to-face mode of

administering the FIM in the patient's home).
3. The 13 items in the FONE FIM Motor scale are appropriate and optimal with respect to scale construction.
4. The 7-point scale used in the FONE FIM is psychometrically optimal.

In Part II of this paper, the Rasch person ability measures obtained from the FONE FIM will be compared with those from the OBS FIM to demonstrate criterion-related validity. If proven to be valid and reliable, the FONE FIM can be used to routinely evaluate the effects of interventions following discharge, and to screen for cases where further interventions may be indicated. Since its feasibility and cost advantages over in-home interviews have already been demonstrated (Smith, 1992), the potential value of the FONE FIM cannot be overestimated.

METHODS

The Setting and Resources

Since 1982, the Northern Alberta Regional Geriatric (NARG) Program has responded to the complex health needs of frail older adults living in Northern Alberta. They present with a wide range of functional disabilities combined with multiple, complex medical and psychosocial problems. Although the most responsible diagnosis may be either a rehabilitation diagnosis such as hip fracture or a medical-surgical diagnosis such as congestive heart failure, there is always a long list of comorbidities for the older adults served by the program.

The NARG Program offers an interdisciplinary team approach to assessment, treatment and rehabilitation in both inpatient and outpatient settings. There are 104 inpatient beds, of which 82 are assessment and rehabilitation beds and 22 are geriatric psychiatry beds. The outpatient settings include two day hospitals and several outpatient clinics. Admission criteria for the inpatient programs include complex medical problems and significant functional impairment. The team focuses on identifying underlying medical problems, stabilizing these interacting conditions, and optimizing function with a view to discharging the older person to an independent living situation whenever possible.

Subjects

The study group consisted of 132 subjects, a subgroup of the 315 patients discharged home from the NARG Program between September and December, 1993 and 1994. Patients were in the study group if they met the following criteria: 1) they were discharged to an independent community living situation, 2) they resided in the greater Edmonton region, 3) they consented to participate, and 4) they agree to a follow-up visit or phone call by signing an informed consent. The reasons for excluding 183 discharged-home patients from the sample included the following: 47 lived outside the City of Edmonton, 38 refused, 29 were readmitted, 20 continued to attend the Geriatric Day Hospital, 28 did not have the discharge FIM, 1 died and 2 went to continuing care institutions within 6 weeks of discharge, 1 was a non-geriatric rehabilitation case, and the remaining 17 were due to unsuccessful follow-ups.

Measure

The FIM (Dodds et al, 1993; Granger et al., 1993; Heinemann et al., 1993; Stineman et al., 1994; Watson et al., 1995) was originally designed as an observational method of assessing an individual's level of functional independence. It was developed by a task force sponsored in 1984 by the American Academy of Physical Medicine and Rehabilitation and the American Congress of Rehabilitation Medicine (Granger et al., 1993; Smith, 1990). Its admission and discharge data formed a basis of the Uniform Data System (UDS) for Medical Rehabilitation. The FONE FIM was designed as the telephone version of the FIM, and has the same 18 items as the FIM (see Table 1). The FIM has been shown to consist of at least 2 dimensions: 13 "motor" and 5 "cognitive" items (Heinemann et al., 1993; Linacre et al., 1994). Each of these items was designed to measure an aspect of functional independence on a 7-point scale: 1 and 2 for "complete dependence", 3-5 for "modified dependence", and 6-7 for "independence" -- 6 for "modified independence", and 7 for "complete independence."

Procedure

The participating patients or their significant others (if the patients were unable to respond) were contacted by telephone five weeks after discharge

Table 1
Rasch Analysis of 18 FONE FIM and OBS FIM Items, 1993 Data (n = 77)

	FONE FIM			OBS FIM		
FIM Items	Item Logits(Error)	INFIT MnSq(Std)	OUTFIT MnSq(Std)	Item Logits(Error)	INFIT MnSq(Std)	OUTFIT MnSq(Std)
Social	-1.10 (.20)	1.2 (0)	1.1 (0)	-1.02 (.21)	1.2 (0)	.9 (0)
Eating	-.87 (.18)	.9 (0)	.8 (0)	-.90 (.19)	0.9 (0)	.8 (0)
Expression	-.64 (.16)	1.4 (1)	1.0 (0)	-.58 (.16)	1.3 (0)	.9 (0)
Grooming	-.47 (.15)	1.3 (1)	1.1 (0)	-.53 (.16)	1.0 (0)	.9 (0)
Dresssing, Upper	-.44 (.15)	.9 (0)	.7 (-1)	-.28 (.14)	1.1 (0)	.8 (0)
Bed Transfer	-.30 (.14)	.5(-2)	.5 (-2)#	-.53 (.16)	0.6(-1)	.5 (-2)
Toileting	-.17 (.13)	1.1 (0)	.7 (-1)	-.11 (.13)	1.1 (0)	.7 (-1)
Dressing, Lower	-.10 (.13)	1.4 (1)	1.0 (0)	-.01 (.13)	1.5 (1)	1.1 (0)
Bladder	-.01 (.12)	1.8 (2)	1.6 (2)*	-.01 (.13)	2.0 (3)	1.8 (2)*
Bowel	-.02 (.12)	.6(-2)	.7(-1)	-.04 (.13)	0.7(-1)	.8 (-1)
Toilet Transfer	-.04 (.12)	.4(-3)	.5 (-2)#	-.06 (.13)	0.6(-2)	.6 (-1)
Memory	.05 (.12)	2.4 (4)	2.5 (4)*	.01 (.13)	1.6 (2)	1.6 (2)*
Comprehension	.23 (.11)	.7(-1)	.7(-1)	.26 (.11)	0.7(-1)	.9 (0)
Problem Solving	.24 (.11)	2.2 (4)	2.2 (4)*	.14 (.12)	1.9 (3)	1.9 (3)*
Walking	.42 (.11)	.8(-1)	.7(-1)	.42 (.11)	0.9 (0)	.9 (0)
Bathing	.80 (.09)	1.0 (0)	1.0 (0)	.78 (.10)	1.0 (0)	1.1 (0)
Tub Transfer	.89 (.09)	.6(-2)	.7(-1)	.93 (.09)	0.7(-1)	.9 (0)
Climbing Stairs	1.54 (.08)	1.8 (4)	1.8 (3)*	1.51 (.08)	1.7 (3)	1.7 (3)*

Model fit Statistics:	Root Mean-Square Std. Error	Adjusted Std. Dev.	Separation	Reliability	# of Strata
Item Statistics:					
FONE FIM:	.13	.61	4.59	.95	6.4
OBS FIM:	.14	.60	4.39	.95	6.2
Person Statistics:					
FONE FIM	.31	.77	2.49	.86	3.7
OBS FIM	.33	.79	2.41	.85	3.5

* Misfitting items; # Muted items

to administer the FONE FIM and arrange for a home visit in the following week. Although the entire 18-item FONE FIM was administered in 1993, only its 13-item motor component was used in 1994 because of a concern about the validity and utility of telephone-based cognitive assessment. During the home visits, however, all 18 FIM items were administered in both 1993 and 1994. This was done based on direct observations of the patients performing various tasks as appropriate, or on patient/proxy reports when direct observations were not feasible or inappropriate.

The 1993 data were collected by 2 RN research assistants and a graduate student in occupational therapy, and the 1994 data were collected by 3 OT students doing their practicums. All raters went through FIM training, which included going through the FIM training guide (*Guide for Use of the Uniform Data Set for Medical Rehabilitation,* 1990), viewing the training videos, and undertaking an interrater exercise. In 1993, all 3 raters independently rated the 18 FIM items based on Sample Case #1 in the Guide. In 1994, the 3 raters were paired, and the FONE FIM and the OBS FIM were administered independently to discharged patients. Their ratings were compared, and reasons for any discrepancies were discussed to achieve a consensus on the definitions of various item scale levels. During the data collection stage, patients were assessed by the same rater on the telephone and at home. In addition, the Mini-Mental State Examination (MMSE) (Folstein et al., 1975) was administered in both years on admission to the rehabilitation program and at the 6-week follow-up.

Data Analysis

Validation of the Motor component of the FONE FIM was based on Rasch rating scale analysis (Wright & Masters, 1982), which generated item difficulty and subject ability measures on a common interval scale (in logits, the natural logarithm of odds). To show that the constructs of the FONE FIM and the OBS FIM closely parallelled each other, separate Rasch analyses were performed on the OBS FIM so that model fit statistics could be generated for each of the 2 modes and compared (Chang & Chan, 1995). Since it was not possible for the same subjects to be assessed by more than one rater, the models used were 2-facet, subjects-by-items, models rather than treating raters as the third facet. The computer programs FACETS (Linacre & Wright, 1993), SYSTAT (Wilkinson, 1990), and SPSS (SPSS Inc., 1996) were used to perform Rasch analyses and other

statistical procedures, and to address the following issues:

Dimensionality. To use the total FIM score or the Rasch score, it is necessary to demonstrate that the items form a single construct or dimension. The Rasch model was first fitted separately to the full, 18-item FONE FIM and the OBS FIM data for the 1993 sample (n=77). The resulting model fit statistics were examined to see if the models exhibited desirable characteristics, such as high item and subject separation and reliability. The fit statistics, the separation index (SI), root mean square calibration error (RMSE), adjusted standard deviation (SD), and reliability index (RI), were related mathematically by:

$$SI = SD / RMSE$$

$$RI = SI^2 / (1 + SI^2).$$

Hence, the separation index has to exceed 2 (or 3) in order to attain the desired level of reliability of at least 0.80 (or 0.90). Misfitting items were then identified by examining the information-weighted and unweighted fit statistics, INFIT and OUTFIT, such as the mean square (MnSq) and its standardized statistic (Std) with a mean of 0 and a standard deviation of 1. Standardized INFIT and OUTFIT statistics were used to detect misfitting items (both ≥ 2) or muted items (both ≤ -2) that reflected dependency in the data. Poorly fitting items were then removed, and fit statistics were further examined after refitting to identify a subset of items that better satisfied the assumption of unidimensionality (Wright, 1996).

Hierarchical Structure. A strength of Rasch analysis is its ability to determine a hierarchical structure for the items in the (unidimensional) model, and order them from the easiest to the most difficult. Such structures were examined for the FONE FIM and the OBS FIM based on Rasch item difficulty estimates. Concordance between the 2 modes was assessed in terms of the intraclass correlation coefficient (ICC). The specific model used was the random-effect model ICC(2,1), in which the total variance was partitioned into effects due to differences between items (or subjects), differences between modes, and error variance (Portney & Watkins, 1993, p. 512). The ICC value higher than 0.75 was considered as an indication of good concordance, and lower than 0.75 as moderate concordance (Portney & Watkins, 1993). In addition, reproducibility of the hierarchical structure was examined by applying Rasch analysis separately to the 1993 and 1994 samples, and the resulting item structures

calibrations, an item-by-item analysis was conducted for each of the 13 items to assess the difference between the estimates obtained from the FONE FIM and the OBS FIM. This difference was divided by the pooled standard error of calibrations to generate a t-statistic, which was plotted against the average of the 2 item-difficulty estimates for ease of interpretation (Altman & Bland, 1983).

Appropriateness of Items. Item-specific fit statistics were used to evaluate the appropriateness of items. Since an important characteristic of a good instrument is its ability to discriminate among the levels of performance, overall item and person separation indices were used as the criteria to determine an optimal item set. In addition, the spacing and gaps between items were examined. This was done by estimating "the number of item strata", defining "distinct strata" as those that were 3 calibration errors apart (Wright & Masters, 1982, p. 92). Mathematically, NDS = [(4 x SI) + 1)] / 3, where NDS = the number of distinct strata, and SI = the separation index. The number of such distinct strata relative to the number of items included in the model was used as an indicator of the efficiency of the instrument constructed (Haley et al., 1994).

Appropriateness of Scales. Alternative scaling strategies were tested by examining the item and person fit statistics associated with such strategies. When examining the most responsible diagnoses, the person separation index was used as the main criterion for choosing among competing scale options due to its importance in scale construction (Green, 1996). The floor and ceiling effects were also examined.

RESULTS

Sample Characteristics

The 132 subjects in our sample were similar to the 315 discharged-home patients during the study period -- in terms of gender (68% vs. 64% female), age (averaged 79 years for both), cognitive status (a mean MMSE score of 25 and 26 for both at admission and follow-up, respectively), lengths of stay (averaged 38 days for both), and use of Home Care services (68% vs. 64%).

On the other hand, a number of differences existed between the 1993 and the 1994 sample. When examining the most responsible diagnosis, there were proportionally more stroke patients (12% vs. 4%), and fewer patients with orthopaedic and cognitive (3% vs. 11%) conditions in the 1993 than the 1994 sample. As well, the 1993 subjects stayed longer on

stayed longer on average (41 vs. 33 days) despite their significantly higher functional status at admission and discharge: the total FIM score averaged 96 and 87 at admission and 107 and 100 at discharge, respectively, in 1993 and 1994. These differences partly reflected the impact of health care restructuring and downsizing that was taking place in Alberta at that time. However, the difference in functional status became insignificant at the 6-week follow-up: the total FIM scores averaged 108 and 107 in 1993 and 1994, respectively. It should be noted that the majority of our discharged patients were cognitively intact, as reflected in their Cognitive scores: 57% and 76% of the patients in the 1993 and the 1994 sample, respectively, scored 33 or higher on the OBS FIM out of the maximum raw total score of 35. Since Rasch models remove observations with extreme scores, this ceiling effect drastically reduced the effective sample size for the Cognitive scale, which is one of the reasons why this paper focuses only on the Motor component of the FONE FIM and the OBS FIM.

In comparison, the characteristics of 72 other patients who were discharged to continuing care institutions during the same period were somewhat different except for gender (67% female): they were slightly older (averaged 82 years), had a significantly lower mean MMSE score of 20, and stayed much longer (averaged 56 days).

Dimensionality

This was first examined by performing separate Rasch analyses of the full, 18 FONE FIM and the OBS FIM items. To ensure the comparability between the 2 modes, only the 1993 sample was used (Table 1). High and comparable levels of fit were demonstrated for both scales: the item separation and reliability indices were 4.59 and 0.95 for the FONE FIM, and 4.39 and 0.95 for the OBS FIM, respectively. Both scales were associated with similar misfitting items: Memory, Problem Solving, Bladder Management, and Climbing Stairs. Transfer-related items (Bed, Toilet, and Tub) showed muted response patterns.

Because of the presence of 2 misfitting Cognitive items in both scales, it was decided to delete all 5 Cognitive items and test the hypothesis that the remaining 13 items of the FONE FIM and the OBS FIM form a single dimension, as shown for the FIM (Heinemann et al., 1993; Linacre et al., 1994). Separate Rasch models were fitted to the Motor component of the FONE FIM and the OBS FIM (n=132), resulting in substantial improvements in model fit (Table 2): the item separation and reliability

indices increased to 6.82 and 0.98 for the FONE FIM and to 6.98 and 0.98 for the OBS FIM; the person separation index also increased from 2.49 to 2.55 for the FONE FIM and from 2.41 to 2.52 for the OBS FIM, and the reliability index increased from 0.86 to 0.87 for the FONE FIM and from 0.85 to 0.86 for the OBS FIM. Bladder Management and Climbing Stairs remained as misfitting items for both the FONE FIM and the OBS FIM, and transfer items (especially for Bed Transfer and Toilet Transfer) were again muted in the Rasch Motor scale.

Bladder Management, which showed an extremely poor fit, was removed with Bowel Management to see if the remaining 11 items would fit the Rasch model better. The result was an improvement in item separation: 7.76 for the FONE FIM and 8.06 for the OBS FIM (Table 3). However, the person separation index declined for both the FONE FIM (from 2.55 to 2.45) and the OBS FIM (from 2.52 to 2.50). The only misfitting item was Climbing Stairs, which may have resulted from the coding of a "1" if the patient was not testable due to risk (Linacre et al., 1994). Since the person separation index was perceived to be a more important criterion than the item separation index from a practical point of view (Green, 1996) -- and it may be desirable from a clinical standpoint to retain Bladder Management and Bowel Management in the instrument -- the more conventional 13-item Motor scale was retained in the remaining investigations.

Hierarchical Structure

Similar hierarchical structures emerged for the FONE FIM and the OBS FIM (Table 2, Figure 1). For both modes, Eating and Grooming were shown to be the easiest items to perform, while Walking, Bathing, Tub Transfer and Climbing Stairs turned out to be the most difficult. The hierarchical orders were less consistent with respect to moderately-difficult items such as Dressing Upper Body, Dressing Lower Body, Bed Transfer, and Toileting. However, this was due partly to the fact that the estimated difficulty levels of these and other items such as Bowel Management, Bladder Management, and Toilet Transfer were all clustered together (Figure 2). Since the number of distinct strata was 9 or 10 for both modes, some of the 13 items inevitably fell in the same stratum. Nevertheless, concordance between the two modes was good: the ICC was 0.99 and 0.91 for the item difficulty and subject ability measures, respectively. Good concordance was also indicated by the fact that all t-

Table 2
Rasch Analysis of 13 Motor Items: FONE FIM vs. OBS FIM

	FONE FIM			OBS FIM		
FIM Items	Item Logits(Error)	INFIT MnSq(Std)	OUTFIT MnSq(Std)	Item Logits(Error)	INFIT MnSq(Std)	OUTFIT MnSq(Std)
Motor Items (n=132)						
Eating	-.90 (.13)	1.6 (2)	1.4 (1)	-1.03 (.14)	1.0 (0)	1.0 (0)
Grooming	-.80 (.13)	1.5 (1)	1.1 (0)	-.81 (.13)	1.0 (0)	.7 (-1)
Toileting	-.53 (.11)	1.1 (0)	.8 (-1)	-.39 (.1)	1.0 (0)	.7 (-1)
Dressing, Upper	-.47 (.11)	.9 (0)	.9 (0)	-.33 (.11)	1.1 (0)	.9 (0)
Bed Transfer	-.29 (.10)	.5 (-3)	.5 (-3)#	-.52 (.11)	.6 (-2)	.5 (-3)#
Dressing, Lower	-.23 (.10)	1.0 (0)	.8 (-1)	-.08 (.10)	1.3 (1)	.9 (0)
Bowel Management	-.18 (.10)	.9 (0)	1.0 (0)	-.28 (.11)	1.0 (0)	1.2 (0)
Bladder Management	-.15 (.10)	2.2 (5)	1.9 (3)*	-.28 (.11)	2.4 (5)	2.0 (4)*
Toilet Transfer	-.10 (.10)	.4 (-4)	.5 (-3)#	-.12 (.10)	.5 (-3)	.7 (-1)
Walking	.28 (.09)	.9 (0)	.8 (-1)	.29 (.09)	1.0 (0)	1.0 (0)
Bathing	.88 (.07)	1.4 (2)	1.3 (1)	.94 (.08)	1.1 (0)	1.1 (0)
Tub Transfer	.94 (.07)	.9 (-1)	1.1 (0)	1.00 (.08)	.9 (0)	1.1 (0)
Climbing Stairs	1.56 (.07)	1.6 (3)	1.7 (3)*	1.60 (.07)	1.4 (2)	1.7 3)*
Model fit Statistics:	Root Mean-Square Std. Error	Adjusted Std. Dev.	Separation	Reliability	# of Strata	
Item Statistics:						
FONE FIM:	.10	.69	6.82	.98	9.4	
OBS FIM:	.10	.72	6.98	.98	9.6	
Person Statistics:						
FONE FIM	.40	1.03	2.55	.87	3.7	
OBS FIM	.42	1.06	2.52	.86	3.7	

* Misfitting items; # Muted items

Table 3
Rasch Analyses of 11 Motor Items: FONE FIM vs. OBS FIM

	FONE FIM			OBS FIM		
FIM Items	Item Logits(Error)	INFIT MnSq(Std)	OUTFIT MnSq(Std)	Item Logits(Error)	INFIT MnSq(Std)	OUTFIT MnSq(Std)
Eating	-1.03 (.14)	1.5 (2)	1.3 (1)	-1.24 (.15)	1.0 (0)	1.1 (0)
Grooming	-.92 (.13)	1.5 (2)	1.1 (0)	-.99 (.14)	1.1 (0)	.8 (-1)
Toileting	-.62 (.12)	1.2 (0)	1.0 (0)	-.50 (.12)	1.1 (0)	.8 (-1)
Dressing, Upper	-.55 (.12)	.9 (0)	.0 (0)	-.44 (.11)	1.2 (0)	.9 (0)
Bed Transfer	-.36 (.11)	.6 (-2)	.5 (-3)#	-.65 (.12)	.6 (-2)	.5 (-2)#
Dressing, Lower	-.29 (.11)	1.1 (0)	.8 (-1)	-.15 (.11)	1.4 (2)	1.0 (0)
Toilet Transfer	-.14 (.10)	.6 (-3)	.6 (-2)#	-.20 (.11)	.6 (-2)	.8 (-1)
Walking	.27 (.09)	.9 (0)	.9 (0)	.27 (.10)	1.1 (0)	1.1 (0)
Bathing	.94 (.08)	1.4 (2)	1.3 (1)	1.02 (.08)	1.2 (1)	1.2 (1)
Tub Transfer	1.00 (.08)	1.0 (0)	1.1 (0)	1.09 (.08)	1.0 (0)	1.2 (1)
Climbing Stairs		1.7 (4)	1.9 (4)*	1.79 (.08)	1.6 (3)	2.0 (5)*
Model fit Statistics:	Root Mean-Square Std. Error	Adjusted Std. Dev.	Separation	Reliability	# of Strata	
Item Statistics:						
FONE FIM:	.11	.83	7.76	.98	10.7	
OBS FIM:	.11	.89	8.06	.98	11.1	
Person Statistics:						
FONE FIM	.45	1.09	2.45	.86	3.6	
OBS FIM	.47	1.17	2.50	.86	3.7	

* Misfitting items; # Muted items

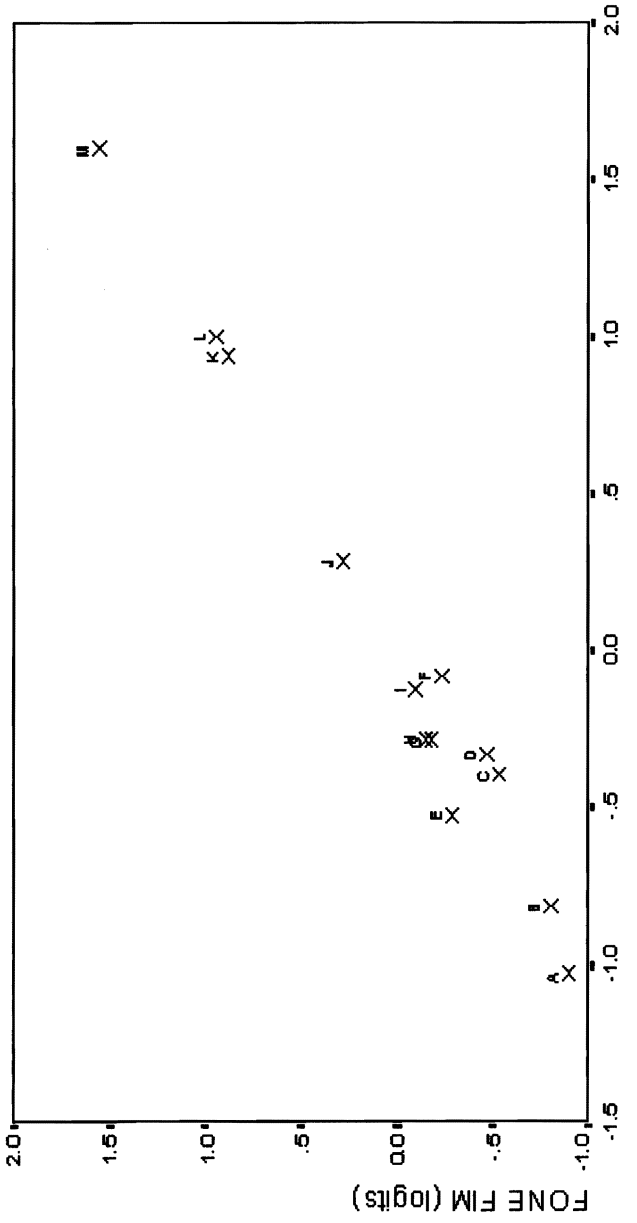


FIGURE 1 Item difficulty estimates FONE FIM vs. OBS FIM.

Measr	-Items	+Subjects (elements	S.1
5	(hard)	(High)	(7)
		***	
4			
		*******	
3		*****	
		*****	
		*******	
		******	
2		****	
		**********	---
	Stairs	******	
		**********	
		**********	
1	TubTr	**********	6
	Bath	**********	
		******	---
		*	
	Walk	****	5
0	ToilTr	*****	---
	Bladder Bowel DressL BedTr	**	4
	DressU	*****	---
	Toilet	*	3
	Groom		---
-1	Eat		2
		*	
		*	---
		*	
-2	(easy)	(low)	(1)
Measr	-Items	* = 1	S.1

FIGURE 2 FONE FIM motor item difficulty and subject ability scores.

statistics for the differences in item difficulty scores were less than 2 in their absolute values (Figure 3).

Reproducibility of the hierarchical structures between the 2 years was further examined by performing separate Rasch analyses for the 1993 and the 1994 sample (Tables 4-5). Climbing Stairs, Tub Transfer, Bathing, and Walking were again the most difficult items in both years for both the FONE FIM and the OBS FIM, and Eating and Grooming remained among the easiest items to perform. There were some differences in the hierarchical structures of the two sets of item difficulty calibrations. For instance, in 1993, the easiest item on the FONE FIM was Eating, but in 1994 it was Grooming. Toileting was a harder item to perform than Bed Transfer in 1993 for the OBS FIM, but not in 1994. The statistical significance of these discrepancies was questionable, however, since many of the 13 items fell within the same strata which numbered no more than 7 distinct ones. Nevertheless, a slightly lower level of concordance for the FONE FIM than the OBS FIM between the 1993 and the 1994 sample was indicated by the ICC values of 0.86 and 0.94, respectively, for the two modes. This was also corroborated by item-by-item analyses, which showed significant t-statistics (>2 in absolute values) in 4 items (Eating, Dressing Upper Body, Bed Transfer, and Climbing Stairs) for the FONE FIM, and in none of the items for the OBS FIM (Figures 4 & 5).

Appropriateness of Items

From a strictly psychometric point of view, the clustering of item difficulty scores contributed to possible item inefficiency or dependence (Haley et al., 1994). The number of distinct item strata associated with the FONE FIM and the OBS FIM was 9.4 and 9.6 for the 13-item Motor scale, respectively. Thus, some of the 13 items in theses Motor scales failed to form distinct strata. The 11-item Motor scale, on the other hand, had an adequate range to form 11 distinct strata (Table 3); however, some of the items such as Bathing and Tub Transfer were clustered together, and others like Tub Transfer and Climbing Stairs were too far apart, for both the FONE FIM and the OBS FIM.

Figure 2 vividly illustrates that the 13-item, FONE FIM Motor scale contained no item with a difficulty level greater than 2 logits, despite a significant proportion of subjects with an estimated Motor function ability in excess of 2 logits. The situation for the OBS FIM was similar. Thus, both the FONE FIM and the OBS FIM could benefit from adding items "more

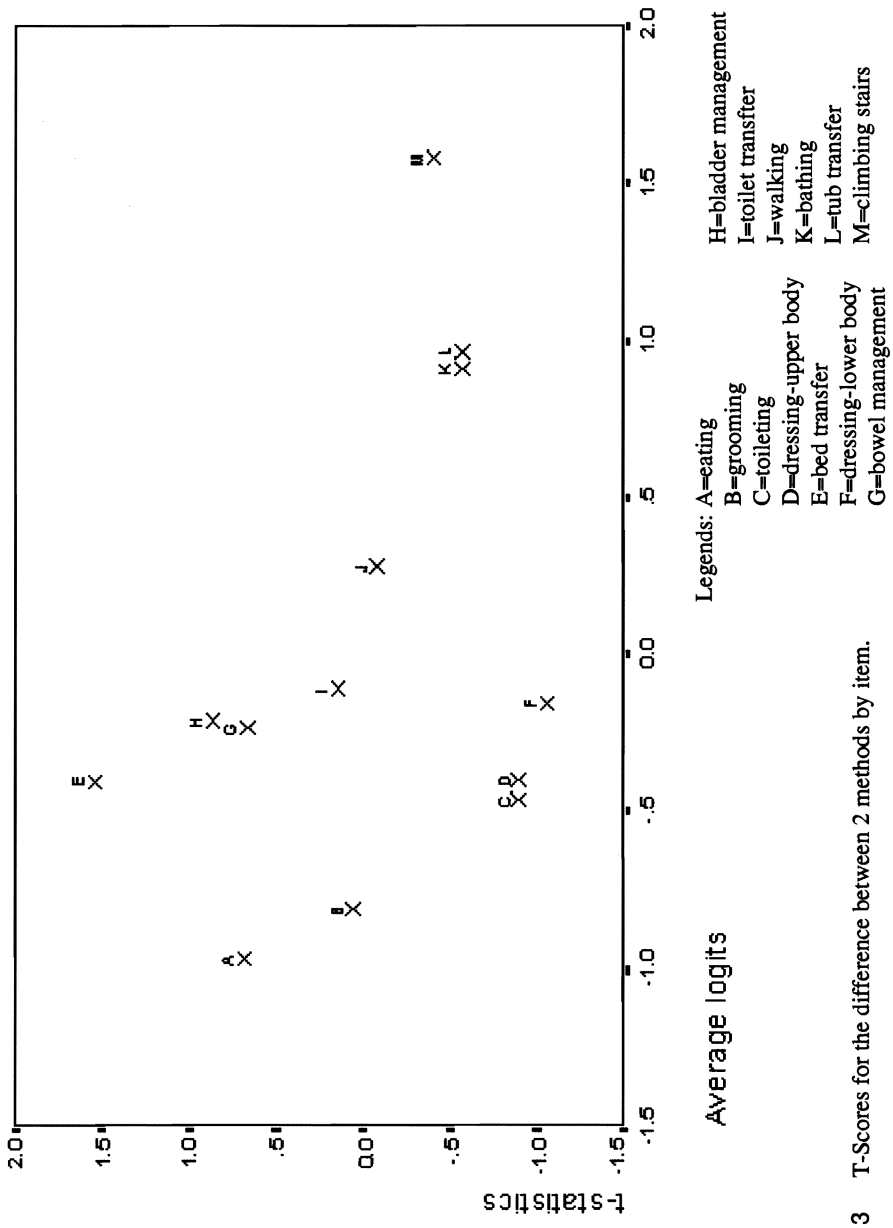


FIGURE 3 T-Scores for the difference between 2 methods by item.

Table 4
Separate Rasch Analyses of FONE FIM Motor Items: 1993 vs. 1994

	1993 Data (n = 77)			1994 Data (n = 55)		
FIM Items	Item Logits(Error)	INFIT MnSq(Std)	OUTFIT MnSq(Std)	Item Logits(Error)	INFIT MnSq(Std)	OUTFIT MnSq(Std)
Eating	-1.26 (.20)	1.2 (0)	.9 (0)	-.55 (.17)	1.5 (1)	1.7(1)
Grooming	-.76 (.17)	1.8 (2)	1.4(1)	-.84 (.19)	1.2 (0)	.6(-1)
Toileting	-.37 (.15)	1.0 (0)	.7(-1)	-.70 (.18)	1.3 (0)	.9 (0)
Dressing, Upper	-.73 (.17)	.9 (0)	.7(-1)	-.22 (.14)	.8 (0)	1.0 (0)
Bed Transfer	-.54 (.16)	.5(-2)	.5 (-2)#	-.08 (.14)	.4 (-2)	.4 (-2)#
Dressing, Lower	-.28 (.15)	1.4 (1)	.9 (0)	-.16 (.14)	.7 (-1)	.7(-2)
Bowel Management	-.16 (.14)	.6(-1)	.8 (0)	-.20 (.14)	1.1 (0)	1.1 (0)
Bladder Management	-.14 (.14)	2.4 (4)	2.0 (3)*	-.18 (.14)	2.1 (3)	1.9 (2)*
Toilet Transfer	-.19 (.15)	.4(-3)	.6 (-2)#	-.01 (.13)	.4 (-2)	.5(-1)
Walking	.44 (.12)	.9 (0)	.8 (-1)	.13 (.13)	.9 (0)	.8 (0)
Bathing	.96 (.11)	1.2 (0)	1.2 (0)	.81 (.11)	1.6 (2)	1.6 (1)
Tub Transfer	1.08 (.11)	.7 (-1)	.9 (0)	.81 (.10)	1.0 (0)	1.3 (1)
Climbing Stairs	1.95 (.10)	1.9 (4)	1.8 (3)*	1.18 (.10)	1.2 (0)	1.3 (1)

Model fit Statistics:	Root Mean-Square Std. Error	Adjusted Std. Dev.	Separation	Reliability	# of Strata
Item Statistics:					
1993 Data	.15	.84	5.72	.97	8.0
1994 Data	.14	.56	3.96	.94	5.6
Person Statistics:	.43	1.04	2.43	.86	3.6
1993 Data	.39	.90	2.34	.85	3.5
1993 Data					

* Misfitting item; # Muted items

Table 5
Separate Rasch Analyses of OBS FIM Motor Items 1993 vs. 1994

	1993 Data (n = 77)			1994 Data (n = 55)		
	Item	INFIT	OUTFIT	Item	INFIT	OUTFIT
FIM Items	Logits(Error)	MnSq(Std)	MnSq(Std)	Logits(Error)	MnSq(Std)	MnSq(Std)
Eating	-1.14 (.20)	1.1 (0)	1.1 (0)	-.91 (.20)	.7 (-1)	.9 (1)
Grooming	-.73 (.17)	1.0 (2)	.8 (0)	-.91 (.20)	1.0 (0)	.6 (-1)
Toileting	-.25 (.14)	.9 (0)	.6 (-1)	-.57 (.17)	1.1 (0)	.8 (0)
Dressing, Upper	-.45 (.15)	1.2 (0)	.8 (0)	-.20 (.15)	1.0 (0)	1.0 (0)
Bed Transfer	-.73 (.17)	.6 (-1)	.5 (-2)	-.30 (.16)	.5 (-1)	.5 (-2)
Dressing, Lower	-.12 (.13)	1.4 (1)	1.0 (0)	-.01 (.15)	1.1 (0)	.8 (0)
Bowel Management	-.15 (.14)	.7 (-1)	.9 (0)	-.35 (.16)	1.5 (1)	1.6 (1)
Bladder Management	-.12 (.13)	2.2 (4)	2.0 (3)*	-.49 (.17)	2.4 (3)	1.8 (2)*
Toilet Transfer	-.20 (.14)	.6 (-2)	.6 (-1)	-.03 (.15)	.5 (-2)	.7 (0)
Walking	.37 (.12)	1.1 (0)	1.0 (0)	.19 (.14)	1.0 (0)	1.0 (0)
Bathing	.81 (.11)	1.2 (0)	1.1 (0)	1.10 (.11)	1.0 (0)	1.0 (0)
Tub Transfer	1.00 (.10)	.8 (-1)	1.0 (0)	1.02 (.11)	1.0 (0)	1.2 (0)
Climbing Stairs	1.70 (.09)	1.7 (3)	1.8 (3)*	1.48 (.11)	1.1 (0)	1.7 (2)

Model fit Statistics:	Root Mean-Square Std. Error	Adjusted Std. Dev.	Separation	Reliability	# of Strata
Item Statistics:					
1993 Data	.14	.74	5.24	.96	7.3
1994 Data	.15	.71	4.63	.96	6.5
Person Statistics:					
1993 Data	.43	1.04	2.43	.86	3.6
1994 Data	.41	1.07	2.64	.87	3.9

* Misfitting items; # Muted items

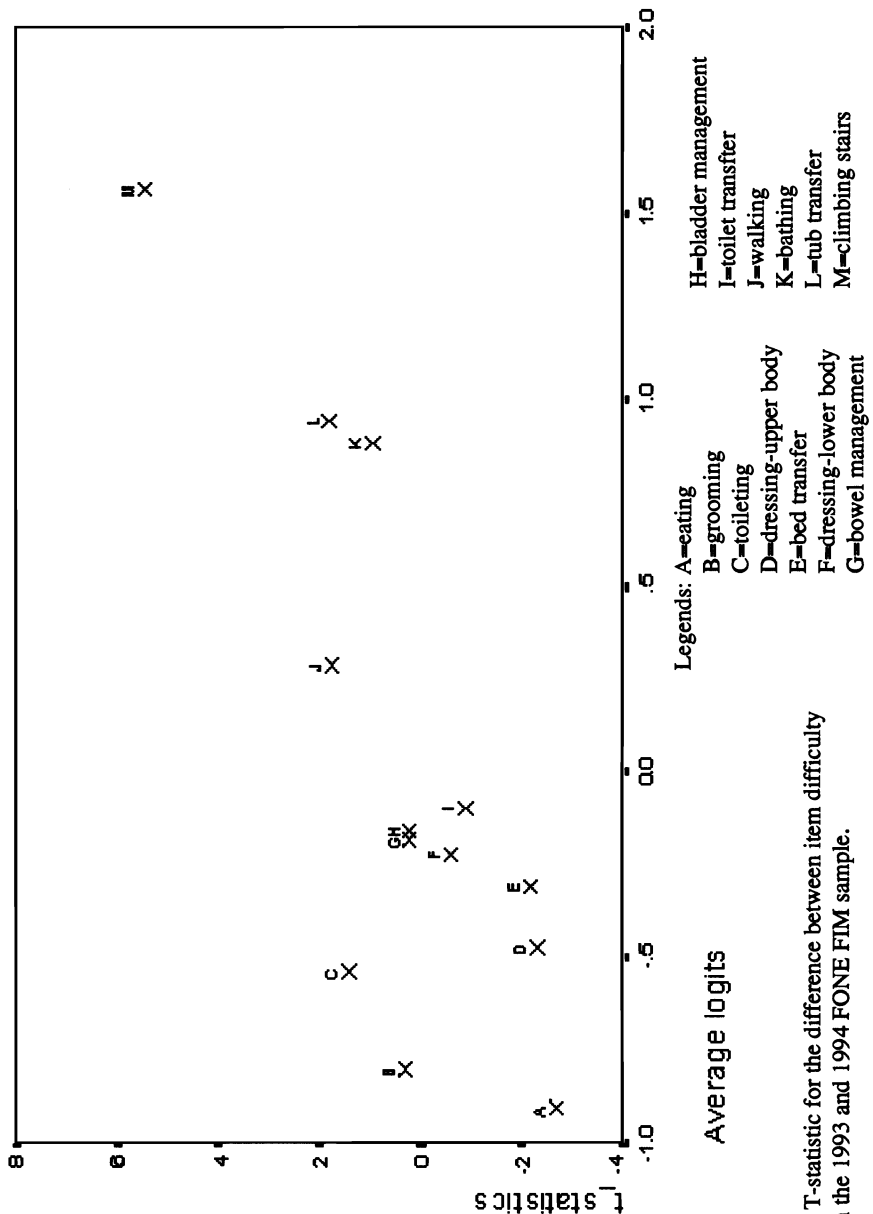


FIGURE 4 T-statistic for the difference between item difficulty estimates from the 1993 and 1994 FONE FIM sample.

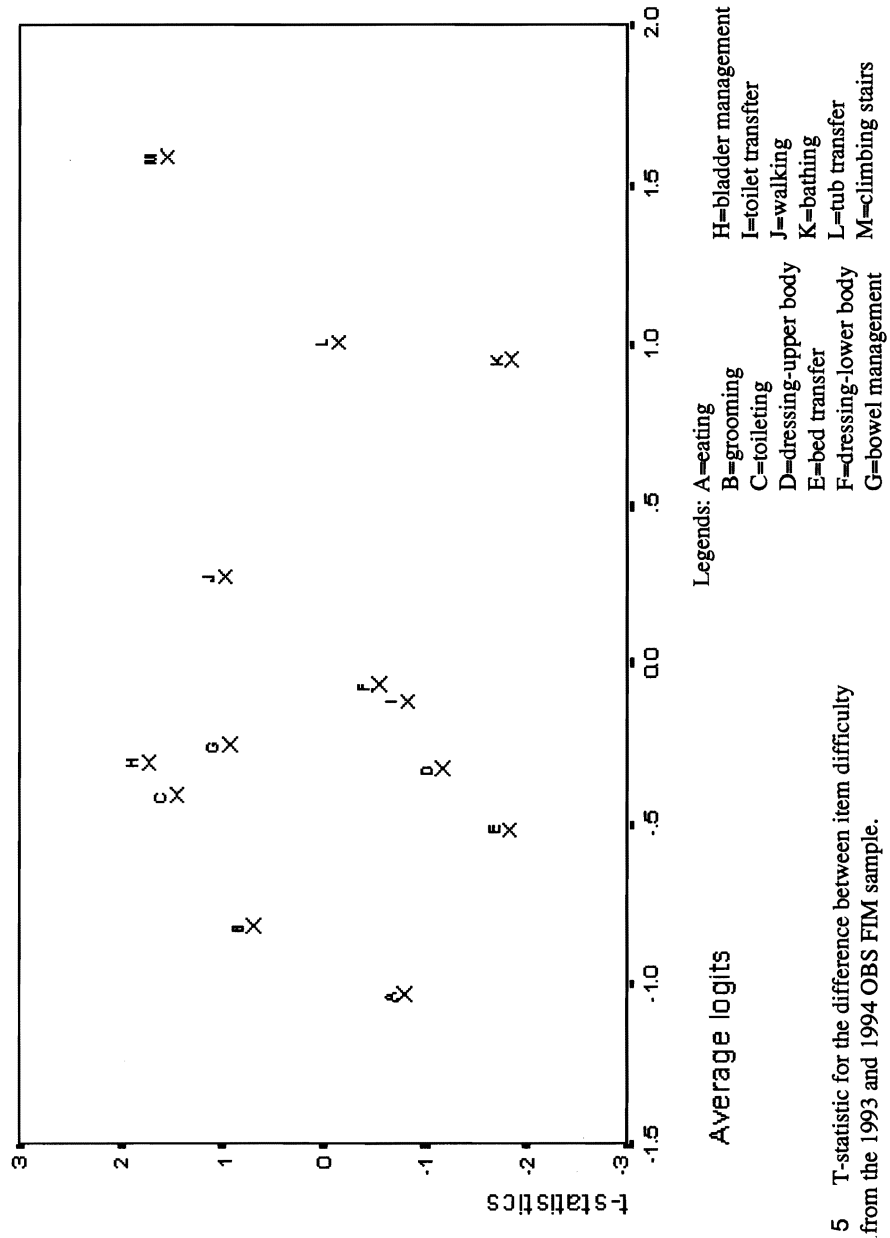


FIGURE 5 T-statistic for the difference between item difficulty estimates from the 1993 and 1994 OBS FIM sample.

difficult" than Climbing Stairs, in order to more adequately assess the levels of functional independence of subjects at the upper end of the Motor ability scale.

Appropriateness of Scales

To assess the appropriateness of the 7-point scale used in the FONE FIM and the OBS FIM, the following scales were constructed by combining some of the levels in the original 7-point scale: (a) a 5-point scale (Levels: 1-2, 3-4, 5, 6, 7), (b) 2 alternative 4-point scales (Levels: 1-2, 3-5, 6, 7) & (Levels: 1-3, 4-5, 6, 7), (c) a 3-point scale (Levels: 1-2, 3-5, 6-7), and (d) a 2-point scale (Levels: 1-5, 6-7). The item and person, separation and reliability indices resulting from fitting them to Rasch models are summarized in Table 6. As the scale levels decreased from 7 to 2, the corresponding separation and reliability indices increased, peaked at the 4-point scales, and then decreased when the levels were further reduced. This is especially the case for the more important criteria of the person (rather than the item) separation and reliability indices. The performance of the two 4-point scales did not differ much, however, in terms of person separation: the first scale did marginally better than the second for the FONE FIM, but did not do as well for the OBS FIM. However, in terms of item separation, the second alternative did better for both modes. Instead of 7 levels, one of these 4-point scales would appear to be a better scaling strategy. Rasch analysis of the first 4-point scale, with a slightly higher person separation index for the FONE FIM, is summarized in Table 7. It should be noted that the item measures were less precise than those derived from the 7-point scale, as reflected in the calibration errors (RMSE=0.15-0.16 vs. 0.10). However, the Rasch model fitted this new scale better than the 7-point scale: Bladder Management remained the only misfitting item in the new 4-point Motor scale.

DISCUSSION

A forceful argument has been made in favour of using Rasch rather than classical analysis to evaluate outcome measures (Wright, 1996). The proposed "individual item editing rule" and a more liberal "hybrid method" further operationalized the procedures for assessing dimensionality and selecting optimal item sets based on maximization of person separation (Green, 1996). This methodology was deployed to evaluate the Motor

Table 6
Item and Person Separation and Reliability Indices for Various Scaling Options

		Separation		Reliability	
Scaling Option	Mode	Item	Person	Item	Person
7-point scale	FONE FIM	6.82	2.55	0.98	0.87
	OBS FIM	6.98	2.52	0.98	0.86
5-point scale	FONE FIM	7.29	2.84	0.98	0.89
	OBS	7.29	2.74	0.98	0.88
4-point scale	FONE FIM	7.62	3.02	0.98	0.90
(1-2,3-5,6,7)	OBS	7.53	2.97	0.98	0.90
4-point scale	FONE FIM	7.68	3.00	0.98	0.90
(1-3,4-5,6,7)	OBS	7.65	3.03	0.98	0.90
3-point scale	FONE FIM	5.58	2.04	0.97	0.81
	OBS	8.11	2.44	0.99	0.86
2-point scale	FONE FIM	4.77	1.19	0.96	0.59
	OBS	4.60	1.34	0.95	0.64

Table 7
Rasch Analyses 4-Point Motor and Cognitive Items: FONE FIM vs. OBS FIM

	FONE FIM			OBS FIM		
FIM Items	Item Logits(Error)	INFIT MnSq(Std)	OUTFIT MnSq(Std)	Item Logits(Error)	INFIT MnSq(Std)	OUTFIT MnSq(Std)
Motor Items (n=132)						
Eating	-1.24 (.17)	1.5 (2)	1.4 (1)	-1.39 (.18)	1.4 (2)	1.3 (1)
Grooming	-1.30 (.18)	1.1 (0)	.8 (0)	-1.26 (.18)	1.0 (0)	.8(0)
Toileting	-.98 (.17)	.9 (0)	.8(-1)	-.87 (.17)	.8 (-1)	.6 (-1)
Dressing, Upper	-.90 (.16)	.9 (0)	.8 (0)	-.76 (.16)	.8 (-1)	.7 (-1)
Bed Transfer	-.50 (.15)	.5(-4)	.5(-3)#	-.76 (.16)	.5 (-4)	.5(-2)#
Dressing, Lower	-.67 (.16)	.9 (0)	.7(-1)	-.51 (.16)	1.1 (0)	.8(-1)
Bowel Management	-.16 (.15)	.9 (0)	1.1 (0)	-.22 (.15)	1.1 (0)	1.4 (2)
Bladder Management	-.50 (.15)	1.8 (4)	1.6 (2)*	-.63 (.16)	1.9 (5)	1.5 (2)*
Toilet Transfer	.06 (.14)	.5(-4)	.6(-2)#	.05 (.15)	.7 (-3)	.8 (-1)
Walking	.71 (.14)	.8 (-1)	.9(0)	.80 (.14)	.7 (-2)	.8(-1)
Bathing	1.01 (.13)	1.4 (3)	1.3 (1)	1.20 (.13)	1.3 (2)	1.2 (1)
Tub Transfer	1.78 (.13)	.7(-2)	.8 (-1)	1.70 (.13)	.7 (-2)	.8(0)
Climbing Stairs	2.68 (.14)	1.3 (2)	1.4 (1)	1.92 (.10)	1.1 (1)	1.2 (-1)

Model fit Statistics:	Root Mean-Square Std. Error	Adjusted Std. Dev.	Separation	Reliability	# of Strata
Item Statistics:					
FONE FIM:	.15	1.16	7.62	.98	10.5
OBS FIM:	.16	1.17	7.53	.98	10.4
Person Statistics:					
FONE FIM	.53	1.60	3.02	.90	4.4
OBS FIM	.54	1.60	2.97	.90	4.3

* Misfitting items; # Muted items

component of the FONE FIM and the OBS FIM, including the determination of their optimal item sets and scale levels.

As expected, the results for the FONE FIM and the OBS FIM largely parallelled each other in terms of the item hierarchical structure, misfitting items, optimal item sets, and optimal scale levels. Reproducibility based on samples from different years was less than optimal especially for the FONE FIM, which, in part, may be attributable to differences in characteristics of the patient populations between the two years. The hierarchical structures and misfitting items (e.g., Bladder Management and Climbing Stairs) in our study were also similar to those of the admission and discharge FIM based on 14,799 records in the Uniform Data System for Medical Rehabilitation (Linacre et al., 1994). These corroborating results reinforce the construct and concurrent validity of the 2 modes, which may be further enhanced by adopting the suggestions made for improving the FIM (Linacre et al., 1994). It would appear that the construct validity of the FIM is unaffected regardless of whether it is used in inpatient or home settings. In view of this, the OBS FIM will be used as an external standard for establishing the criterion-based validity for the FONE FIM in Part II of the paper.

It should be noted, however, that both the FONE FIM and the OBS FIM became "too easy" due to patients' improved functional status. This problem was manifested in a large ceiling effect for the Cognitive scale, and in the absence of items harder than Climbing Stairs in the Motor scale. To address this problem, it is necessary to augment the FIM with "more difficult" items such as those related to instrumental activities of daily living (Fillenbaum, 1985; Siu et al., 1990), or to develop a new scale by co-calibrating, for example, the FIM and the Physical Functioning component of the SF-36 based on reconstructed scale levels (Heinemann et al., 1996).

That a 4-point scale turned out to be an optimal scaling strategy is of some historical interest. The FIM originally had 4 levels corresponding to the first 4-point scale described earlier, but was increased to 7 levels in 1987 on the recommendation of clinicians to enhance sensitivity (Hamilton et al., 1987; Hamilton et al., 1994). Although our results indicated that the original 4-point scale may be superior to the current 7-point scale from a strictly psychometric standpoint, it may still be desirable to retain the 7-point scale from a clinical standpoint. The inclusion of more explicit criteria in the UDS guidelines could improve the clinicians' ability to discriminate between the 2-4 ratings of dependence. Further research is required to address the issue of optimal scale levels from both

psychometric and clinical perspectives.

CONCLUSION

Is the FONE FIM as valid as the face-to-face mode? The answer appears to be a qualified "yes" for its Motor component, in view of the similar item hierarchical structures, misfitting items, optimal item sets, and optimal scale levels. However, there is some evidence indicating that reproducibility may be a problem especially for the FONE FIM . Thus, more detailed comparisons are required to assess the degrees of concordance and ability to substitute between the two modes. The results of such an investigation will be reported in Part II of this paper.

REFERENCES

Altman, D.G., & Bland, J.M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician, 32,* 307-317.

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

Dodds, T.A., Martin, D.P., Stolov, W.C., & Deyo, R.A. (1993). A validation of the Functional Independence Measurement and its performance among rehabilitation inpatients. *Archives of Physical Medicine & Rehabilitation, 74,* 531-536.

Fillenbaum, G.G. (1985). Screening the elderly. A brief instrumental activities of daily living measure. *Journal of the American Geriatrics Society, 33,* 698-706.

Folstein, M.F., Folstein, S.E., & McHugh, P.R. (1975). "Mini-Mental State" A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189-198.

Granger, C.V., Cotter, A.C., Hamilton, B.B., Fiedler, R.C., & Hens, M.M. (1990). Functional assessment scales: a study of persons with multiple sclerosis. *Archives of Physical Medicine & Rehabilitation, 71,* 870-875.

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 & Rehabilitation, 72,* 84-89.

Green K. (1996). Dimensional analyses of complex data. *Structure Equation Modeling, 3(1),* 50-61.

Guide for Use of the Uniform Data Set for Medical Rehabilitation (1990). Buffalo, NY: State University of New York at Buffalo.

Guyatt, G.H., Feeny, D.H., & Patrick, D.L. (1993). Measuring health-related quality of life. *Annals of Internal Medicine, 118(8),* 622-629.

Haley, S.M., McHorney, C.A., & Ware, J.E., Jr. (1994). Evaluation of the MOS

ibility of the Rasch item scale. *Journal of Clinical Epidemiology, 47(6),* 671-84.

Hamilton, B.B., Granger, C.V., Sherwin F.S., Zielezny, M., & Tashman J.S. (1987). A uniform national data system for medical rehabilitation. In M.J. Fuhrer (Ed.), *Rehabilitation Outcomes: Analysis and Measurement* (pp. 137-147). Baltimore, MD: Brookes.

Hamilton, B.B., Laughlin, J.A., Fiedler, R.C., & Granger, C.V. (1994). Interrater reliability of the 7-level Functional Independence Measure (FIM). *Scandinavian Journal of Rehabilitation Medicine, 26(3),* 115-119.

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 & Rehabilitation, 74,* 566-573.

Heinemann, A.W., Segal, M., Schall, R.R., & Wright B.D. (1996, May). *Extending the range of the Functional Independence Measure with SF-36 items.* Paper presented at the First International Outcome Measurement Conference, Chicago, IL.

Kidd, D., Stewart, G., Baldry, J., Johnson, J., Rossiter, D., Petruckevitch, A., & Thompson, A.J. (1995). The Functional Independence Measure: A comparative validity and reliability study. *Disability and Rehabilitation, 17(1),* 10-14.

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

Linacre, J.M., & Wright, B.D. (1993). *FACETS, Many-Facet Rasch Analysis with FACFORM, Data Formatter.* Chicago, IL: MESA Press.

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

Portney, L.G., & Watkins, M.P. (1993). *Foundations of Clinical Research. Applications to Practice.* Norwalk, CT: Appleton & Lange.

Siu, A.L., Reuben, D., & Hays, R.D. (1990). Hierarchical measures of physical function in ambulatory geriatrics. *Journal of the American Geriatrics Society, 38,* 1113-1119.

Smith, P. (1992). Collecting followup data. *UDS Update, August,* 1-2.

Smith, P., Hamilton, B.B., & Granger, C.V. (1990). *The FONE FIM.* Buffalo, NY: Research Foundation of the SUNY.

Smith, P.M., Illig, S.B., Fielder, R.C., Hamilton, B.B., & Ottenbacher, K.J. (1996). Intermodal agreement of follow-up telephone functional assessment using the Functional Independence Measure in patients with stroke. *Archives of Physical Medicine & Rehabilitation, 77(5),* 431-435.

SPSS Inc. (1996). *SPSS 7.0-7.5.* Chicago, IL 60611.

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.

Watson, A.H., Kanny, E.M., White, D.M., & Anson, D.K. (1995). Use of standardized activities of daily living rating scales in spinal cord injury and disease services. *American Journal of Occupational Therapy, 49(3),* 229-234.

Weinberger, M., Oddone, E.Z., Samsa, G.P., & Landsman, P.B. (1996). Are health-related quality-of-life measures affected by the mode of administration? *Journal of Clinical Epidemiology, 49(2),* 135-140.

Wilkinson, L. (1990). *SYSTAT.* Evanston, IL: SYSTAT, Inc.

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

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

JOURNAL OF OUTCOME MEASUREMENT, *1*(3), 219 - 238

Scoring and Analysis of Performance Examinations: A Comparison of Methods and Interpretations

Mary E. Lunz
American Society of Clinical Pathologists

Randall E. Schumacker
University of North Texas

The purpose of this study was to compare the results and interpretation of the data from a performance examination when four methods of analysis are used. Methods are 1) traditional summary statistics, 2) inter-judge correlations, 3) generalizability theory, and 4) the multi-facet Rasch model. Results indicated that similar sources of variance were identified using each method; however, the multi-facet Rasch model is the only method that linearized the scores and accounts for differences in the particular examination challenged by a candidate before ability estimates are calculated.

Requests for reprints should be sent to Mary E. Lunz, Board of Registry, American Society of Clinical Pathologists, 2100 W. Harrison St., Chicago, IL 60612-0277.

The purpose of a performance examination is to infer candidate abilities beyond the particular sample of tasks, projects and judges on the examination. Whether the goal is to make reproducible pass/fail decisions or to position candidates according to their demonstrated ability, the performance examination must measure candidate ability as accurately as possible. How the performance examination is structured, as well as, the scoring and analysis methods used, have a significant impact on the interpretation and reliability of the examination score and outcome.

To better understand the structure of performance examinations, it is useful to break the examination down into its facets, so the influence of each facet on the score can 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 candidate may be rated by severe judges on difficult projects, while another candidate may be rated by more lenient judges on projects of moderate difficulty. The relevance of these examination facets controls the "meaning" of the score. Reliability implies the reproducibility of the score, and is influenced by the structure and relevance of the examination, as well as, the precision with which candidates are evaluated. When the structure of the examination differs among candidates, the meaning of the score is neither generalizable nor reproducible.

FACETS OF A PERFORMANCE EXAMINATION

There are typically four separate facets in a performance examination. The first facet is **candidate ability,** which encompasses the knowledge and skill possessed by the candidate with regard to the problem, task or product measured by the performance examination. It is expected that candidates will vary in their ability. The goal of the examination is to differentiate among candidates reliably.

The second facet is the **projects** or in this example **topic**. Some projects have detailed specifications that are comparable across candidate performances. Examples are medical cases, essay prompts, science, or laboratory projects. The requirements are described to candidates who then perform to the best of their ability. Other performance examinations, allow candidates to select a sample of their own work. In medicine, candidates may select cases from their medical practice to present in an oral examination. Portfolios may be developed in art or writing. The

performances usually cover specific content specifications so that general areas of knowledge and skill are represented. Sometimes cases are structured and all candidates are challenged by the same group of cases. In the example presented, there are three topics labelled 1, 2, 3, that covered the pertinent areas of the medical specialty.

The third facet is the **judge**. Judges are essential for performance examinations; however, judges have unique physical and mental characteristics, as well as, unique reactions to the examination, all of which influence ratings. Some judges give consistently lower ratings across candidates while others tend to be more generous. Training focuses and directs a judge's attention, but it is usually unable to alter permanently the knowledge and skill that has developed over a lifetime (Stahl and Lunz, 1994).

The fourth facet is the **tasks** or rating dimensions associated with each project or case. Considerations for this facet include: (1) the number of tasks rated, (2) the extent of detail in the definition of the tasks, and (3) the relevance of the tasks to the cases or projects. Tasks may be fairly objective, such as using correct punctuation, or may be more subjective such as ethical standards for medical treatment. Tasks must be carefully delineated. The tasks in the example presented are the ability to: 1) recall factual information, 2) interpret data, 3) solve clinical problems presented in the topic areas.

The definition of the **rating scale** provides a "disciplined dialogue" which encourages raters to assign specified meaning to each category on the scale. Rating scales may have as few as two categories (0/1) or an infinite number (0/∞). Usually, each category on the scale has a specific definition. The definitions of the rating scale categories are important, because they influence how judges use the scale. The measurement distance between categories impacts the ratings given to the candidate. For example, there is a great distance between "unacceptable" and "excellent," so logical categories between these extremes are often inserted, (e.g. marginal, acceptable).

The structuring of these facets, and/or others, plus the rating scale, produces the examination design. The design of the examination affects the meaning or interpretation of the scores. This fact has been alluded to, but never clearly stated in the literature. For example, LeMahieu, Gitomer, Eresh, (1995) stated that instability in estimates of student performance is introduced by rater judgements and variability in the projects, whether the

projects are defined by the candidate or the examination developer. Bond (1996) states that one problem is the sheer difficulty in scoring complex performance assessments reliably and validly. Efforts have been made to standardize the scoring by training raters, defining inferences, and/or sampling tasks or projects. However, these efforts have not been completely successful. As Guion (1996) points out, performance scores should permit fair, meaningful comparisons and validity reducing error should be minimal. This is most important when the consequences of assessment are considered (Messick, 1995). The issue is that performance based on testing is evolving and changes require that we modify how we think about scoring performance examinations to minimize error and improve precision and reliability.

STANDARDIZATION: HOW MUCH IS ENOUGH?

The issue of what and how much to standardize in a performance examination determines the examination design. In order for any consistency to be achieved, the rating scale must be standardized. All judges must use the same rating scale, understand the definitions assigned to the rating scale categories, and be willing to use the scale when rating candidate performance. The rating scale is the foundation for scoring candidates.

The tasks or rating dimensions, the actions required of the candidate, are also relatively easy to standardize. Pertinent tasks are defined explicitly and agreed upon by the judges. If three tasks are defined, then all judges must be willing to assess candidates on the three tasks. If six tasks are defined, then all judges must use the six tasks to rate candidate performance.

The next tier to standardize is the projects, cases, or topics, etc. The most standardization occurs when project material is prepared by the examination developer or committee for all judges to use. All judges then use the same standardized project materials and defined questions and concepts as the basis for examining the candidate. Examples are structured protocols for an oral exam or standardized essay prompts. Somewhat less standardization occurs when the candidate produces projects using detailed specifications and requirements. Examples are essays on specified topics or laboratory slides using specific tissue/stain combinations. Even less standardization occurs when candidates or examiners select the material to be presented. For example, students create portfolios of their art work, or examiners select medical cases from their

practice to present at an oral examination. Definition of the case topic areas produces some standardization and is loosely equivalent to following content guidelines. As the standardization of the examination materials decreases, the need for standardization of the tasks increases to insure consistent measurement among candidates.

Standardizing judges through extensive training and re-training is very difficult, but inter-judge reliability coefficients have traditionally been associated with performance examination reliability. The literature suggests that judge training is not completely successful, at least based on interjudge correlational analysis. In fact, even if judges do correlate perfectly, there is still no guarantee that they would rate candidate performance comparably (Lunz, Stahl, and Wright, 1994). It is interesting to note that standardizing judges, the success of which is measured by correlational analysis, which is the least controllable of all methods of achieving reliability, has been used most frequently for establishing reliability (LeMahiew, Gitomer, and Eresh, 1995).

EXAMINATION RELIABILITY

Examination reliability is ultimately controlled by the design of the examination. The number and consistency of the ratings given to candidate performances influences the precision of the candidate ability estimate. The precision with which candidate performance is measured is controlled, in large part by the number of ratings given to a candidate. Performance examinations have traditionally included a low number of ratings. Thus, the face validity of the examination may be high, but the confidence in the pass/fail decision or candidate placement is low, because the error of measurement is high. Increasing the number of ratings when designing performance examinations decreases the error of measurement, increases the precision of the examination and the confidence in the pass/fail decision. The ratings, however, must be meaningful and consistent within the context and standards of the examination. Traditional methods are often used to indicate inter-rater consistency. Generalizability theory methods can be used to indicate the number of raters and tasks needed to yield a dependable score. Item Response theory methods calculate the measurement error associated with each ability estimate so that the confidence in pass/fail decisions can be determined. The method a psychometrician uses may convey a different meaning and score interpretation.

METHODS OF ANALYSIS FOR PERFORMANCE EXAMINATIONS

Scoring and analysis methods from several social science and educational perspectives have been suggested for analyzing performance examinations. Typically, candidates take different **forms** of the performance examination because they have different sets of judges, submit or challenge different projects at different times during the day or the year. The following analysis methods have been suggested for performance examinations: traditional summary statistics, inter-judge reliability, generalizability-theory, and the multi-facet Rasch model.

Traditional summary statistics include means, standard deviations and reliability coefficients, e.g. Cronbach's Alpha. Cronbach's coefficient Alpha can be viewed as an estimate of the squared correlation between the observed score and the true score. When there is a high correlation between the observed and true score, it suggests low measurement error and high reliability. These traditional summary statistics are familiar to candidates and test developers.

Inter-judge correlation coefficients are frequently used to ascertain the reliability of performance examinations. This is based on the premise that if judges' ratings correlate well, then a random sample of judges can be assigned to any candidate, and if the two randomly selected judges agree, the score must be true. Many researchers have found that inter-judge reliability is not stable. Blak (1985) did a systematic test-retest study in which the same 16 judges graded the same 105 essays on two different occasions. Interjudge reliability was defined as the relationship among the sets for different raters, with the hypothesis that the scores of all raters correlate perfectly. Correlational estimates ranged between .55 and .84 indicating rater-specific idiosyncratic interpretations. Blak's conclusion was that different raters must be viewed as different measuring instruments and that the median values of the correlations, .45 and .48, should serve as the reliability estimate.

Generalizability theory (G-theory) methods partition the sources of error variance for each facet of the examination. The theory is that measurement error introduces significant variability within a facet. The goal is to improve the measurement design so that precision and reliability will improve for the **next** examination. If the precision is not satisfactory, the study is re-designed to increase reliability. The candidates are considered the "objects of measurement", not a facet in the analysis, so

their scores stand regardless of the results of the G-theory study. Thus, if judges or cases are found to be significantly different, this fact is identified, but the impact on the candidates' scores during a particular examination is not corrected. Recent developments in G-theory methods now permit an identification of which elements in a facet contribute to significant variability, hence measurement error (Marcoulides and Drezner, 1997). A direct method for estimation of true scores in G-theory which does not rely on estimates of rater severity has also been presented (Longford, 1994). Although rater severity estimates are not considered to be necessary for estimation of true scores, they are still useful for identifying erratic raters.

The multi-facet Rasch analysis, extends the basic Rasch model (Rasch, 1960/1980) so that facets for task and item difficulty, and judge severity are added to the equation (Linacre, 1989). This controls the impact of error variance within each facet on the candidate's ability estimate. The probability of a satisfactory performance is a function of the difference between the candidate's ability and the task difficulty, after adjustment for the severity of the judge(s) and the difficulty of the cases/ projects with consideration for usage of the rating scale. If the candidate's ability is higher than the difficulty of the tasks after adjustment for the case difficulty and the judge severity, then the probability of a satisfactory performance is greater than 50%. Conversely, if the task difficulty after adjustment for judge severity, is greater than the ability of the candidate, the probability of achieving a satisfactory performance is less than 50% . This is modelled:

$$\log(P_{nmjik} / P_{nmji(k-1)}) = (B_n - T_m - C_j - D_i - R_k)$$

where: P_{nmjik} is the probability of being rated in category k

$P_{nmji(k-1)}$ is the probability of being rated in category k-1

B_n is the ability of the candidate n

T_m is the difficulty of the task m

C_j is the severity of the judge j

D_i is the difficulty of the project i

R_k is the difficulty of being rated in category k rather than category k-1

The ordering of the candidates, tasks, judges, and projects on a log-

linear scale provides a frame of reference for understanding the relationship of the facets of the performance examination. It makes it possible to observe estimated candidate ability from highest to lowest, estimated task difficulty from most to least difficult, estimated judge severity from most to least severe, and estimated project difficulty from most to least difficult.

DATA

The data from a medical specialty certification examination were analyzed using the four methods to ascertain the similarities and differences of the interpretations that would be drawn from each analysis. The candidates met specific requirements for education and experience to gain entry into the certification process. The data set had four facets: (1) candidates; (2) judges; (3) topics; and (4) tasks. A total of 74 candidates challenged this examination. Three tasks were used as the basis for rating: (1) recall of factual information; (2) interpretation of data; and (3) clinical problem solving. A five point rating scale was used in which A (4 points) was excellent, B (3 points) was above average, C (2 points) was average, D (1 point) was below average and F (0 points) was failing. The 31 judges were qualified experts who were trained in the examination process. Three pairs of randomly assigned judges rated a candidate, for a total of six judges per candidate. Pairs of judges rotated. Therefore, the inter-judge correlations were based on ratings given to the random sample of candidates that pairs of judges had in common. In this example, pairs of judges had 0 to 9 candidates in common. Pairs of judges who had only 1 candidate in common could not be correlated, and there were many pairs of judges who had 0 candidates in common (missing data).

The judges enjoyed a great deal of flexibility in how they examined candidates on each topic; however, they all rated candidates on the same three tasks. Each judge rated approximately 14 candidates during the examination administration. The judges rotated so that all judges had all topics, and some candidates in common during the course of the examination. Thus, judges had all topics, all tasks and some candidates (range = 0-9) in common. The same rating scale was used by all judges for all tasks, for all candidates.

Traditional summary statistics and Cronbach's Alpha reliability estimates were calculated. Inter-judge correlation coefficients were calculated using the Pearson Product Moment coefficient formula.

Sources of variance in the facets were computed using the principles of G-theory and the GENOVA program (Crick and Brennan, 1983). The multi-facet Rasch analysis was completed using FACETS (Linacre, 1990), a computer program for the multi-facet analysis of performance examinations.

RESULTS

Table 1 shows the mean, standard deviation and the Cronbach's Alpha Reliability estimate for candidates, tasks, and topics. Total raw scores for candidates, tasks and topics were used to calculate Cronbach's Alphas. Overall, the performance examination produced a reasonable set of candidate total scores (Cronbach's Alpha = .91). When tasks, and topics were analyzed, more error produced varying reliability coefficients (.70 - .93). Raw scores were used in this analysis, so potential differences in the examinations challenged by candidates were not taken into account.

The Pearson Product Moment correlation coefficients for judges' ratings of candidates ranged from -1.00 to 1.00 depending upon the number of candidates pairs of judges had in common, the topic rated and the perception of the candidate's performance. Due to the rotational system, the number of common candidates between any two judges varied. Thus, a number of factors influenced the inter-judge correlation coefficients, and the range of correlations among judges precipitates questions about the comparability of examinations when candidates are assessed by different judges.

Table 2 shows the results of the generalizability theory analysis. Ranking of persons by topics indicated 31% of the error variance implying that some persons were better in some topics than others, as expected in the examination design. Judges were also not consistent in rating persons within a task (15% error variance). More important, judges ratings of a person on a task within a topic produced a significant amount of error variance (25%).

Tables 3, 4, 5 show the results of the multi-facet analysis for topic and task difficulty, and judge severity respectively. Figure 1 shows the overall placements of the facets of this examination. There was little difference between Topics 2 and 3 with Topic 1 being the most difficult. The task of Interpretation was the most difficult, and Recall was the least difficult. Judges ranged in severity from 1.36 (most severe) to -1.21 logits (least severe) with a .71 logit standard deviation around a mean of zero.

Table 1
Traditional Examination Statistics

FACETS	Cronbach's Alpha	Score Mean	Score SD
Candidates Total Score based on ratings of judges on all topics and tasks			
	.91	51.9	9.1
Tasks including all ratings given to all candidates by all judges			
Recall	.88	17.9	2.9
Interpretation	.93	16.8	3.4
Problem Solving	.90	17.1	3.2
Topics including all ratings given to all candidates by all judges			
One	.70	17.5	3.7
Two	.75	16.9	4.4
Three	.73	17.5	3.9

Ratings = 4 (excellent) to 0 (failing)

Table 2
Generalizability Analysis

Source of Variability	Estimated Variance Components	Percentage (%)
Person (P)	.1579199	23.30
Topic (T)	0	0
Judge: Topic (J:P)	0	0
I (Tasks)	.0096878	1.48
PT	.2069964	30.58
PJ:T	.1012692	14.92
PI	.0038257	0.59
TI	.0025196	0.29
JJ:T	0	0
PTI	.0245074	3.54
PIJ:T	.1709758	25.26
	.677	100%

Table 3
Topics in Difficulty Order

N topics	Obsvd Score	Obsvd Count	Obsvd Average	Fair Average	Calibrated Difficulty	Model S.E.	Infit MnSq	Outfit MnSq
					Most Difficult			
1 TOPIC 1	1251	444	2.8	2.2	.19	.08	1.0	1.0
2 TOPIC 2	1295	444	2.9	2.3	-.09	.08	1.1	1.1
3 TOPIC 3	1296	444	2.9	2.3	-.10	.08	0.9	0.9
					Least Difficult			
Mean (Count: 3)	1280.	444.	2.9	2.3	.00	.08	1.0	1.0
S.D.	21.	0.	0.0	0.1	.13	.00	0.1	0.1

RMSE .08 Adj S.D. .11 Separation 1.34 Reliability .64

Table 4
Tasks in Difficulty Order

N task	Obsvd Score	Obsvd Count	Obsvd Average	Fair Average	Calibrated Difficulty	Model S.E.	Infit MnSq	Outfit MnSq
					Most Difficult			
2 INTERPRETATION	1240	444	2.8	2.1	.25	.08	1.0	1.0
3 PROBLEM SOLVE	1271	444	2.9	2.2	.07	.08	0.9	1.0
1 RECALL	1331	444	3.0	2.4	-.32	.08	1.0	1.0
					Least Difficult			
Mean (Count: 3)	1280.	444.	2.9	2.3	.00	.08	1.0	1.0
S.D.	37.	0.	0.1	0.1	.24	.00	0.0	0.0
RMSE .08 Adj S.D. .23		Separation	2.84	Reliability	.89			

Table 5
Judges in Severity Order

JUDGES	Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrge	Calibrated Severity	Model S.E.	Infit MnSq	Outfit MnSq
					Most Severe			
34	116	42	2.8	1.5	1.36	.25	1.1	1.1
03	77	33	2.3	1.5	1.33	.26	0.9	1.0
27	113	45	2.5	1.8	.90	.23	1.4	1.3
29	109	42	2.6	1.8	.90	.24	0.4	0.4
18	109	45	2.4	1.8	.89	.22	0.6	0.7
09	125	45	2.8	1.8	.85	.24	0.7	0.8
08	118	45	2.6	2.0	.59	.23	1.2	1.2
11	114	45	2.5	2.0	.59	.23	1.4	1.5
16	117	42	2.8	2.0	.51	.25	0.8	0.9
25	123	48	2.6	2.0	.48	.22	1.0	0.9
12	130	45	2.9	2.1	.27	.25	1.4	1.4
01	119	42	2.8	2.1	.27	.26	0.8	0.8
15	127	45	2.8	2.2	.25	.24	0.6	0.6
10	127	45	2.8	2.2	.24	.25	1.3	1.3
20	130	45	2.9	2.3	.00	.25	1.0	1.0
33	127	42	3.0	2.3	-.16	.27	0.6	0.6
28	122	42	2.9	2.4	-.17	.26	1.5	1.6
21	129	45	2.9	2.4	-.23	.25	0.7	0.7
23	137	45	3.0	2.4	-.29	.26	0.5	0.5
13	139	45	3.1	2.4	-.35	.26	1.4	1.3
26	142	45	3.2	2.5	-.44	.27	0.5	0.5
32	135	45	3.0	2.5	-.47	.26	0.9	1.1
07	138	45	3.1	2.5	-.50	.26	1.2	1.2
30	128	42	3.0	2.5	-.56	.27	0.7	0.8
31	139	45	3.1	2.5	-.58	.26	1.1	1.1
06	145	48	3.0	2.5	-.62	.25	1.4	1.3
24	137	45	3.0	2.6	-.81	.26	0.9	0.9
22	138	42	3.3	2.6	-.90	.29	0.9	1.0
19	85	27	3.1	2.7	-.94	.34	1.0	1.0
14	142	45	3.2	2.8	-1.20	.27	1.0	1.0
04	105	30	3.5	2.8	-1.21	.39	1.9	2.2
					Least Severe			
Mean N=31	124.	43.	2.9	2.2	.00	.26	1.0	1.0
S.D.	15.	4.	0.3	0.3	.71	.03	0.4	0.4

RMSE .26 Adj S.D. .66 Separation 2.51 Reliability .86

Note. All judges rated all topics on all tasks for some candidates.

Logit	+Candidate	-Topic	Judge	-Task	S.1
5	*				(4)
	*				
	*				
4					
	*				
	**				
	*				
	******				---
	*				
3	**				
	****				
	*				
	****				
	**				
	***				
	*****				
2	*				3
	**				
	**				
	****				
	*				
	***		**		
1	**				
	***		****		
	*				
	*		***		
	*****		*		
	*		****	Recall	
	*****	Topic 1			
0	***		*	Interpertation	
	*	Topic 2 Topic 3	**		
	*		***	Problem Solving	
			***		
			***		2
	*				
	*		**		
-1		*	*		
			**		
	*				---
-2					(0)
Logit	* = 1	-subject	* = 1	-task	S.1

Candidate Separation Reliability .92
*Represents one candidate or one judge

FIGURE 1 Linear relationship of examination FACETS.

Significant differences in judge severity, task and topic difficulty are shown. Candidate ability estimates ranged from 4.98 (highest) to -1.30 (lowest) logits. These ability estimates accounted for the particular judges encountered by the candidate. Candidate Separation Reliability was .92. This indicates that the examination successfully differentiated among candidate performance after the characteristics of the particular examination challenged are accounted for. The multi-facet Rasch model is the only analysis method that accounts for examination characteristics before an ability estimate is calculated.

Not accounting for differences in the severity of the judges can have a major impact on the examination outcomes for some candidates. Table 6 shows two candidates who earned the same raw scores, but substantially different ability estimates when judge severity was taken into account. Candidate 411 had a significantly higher probability of success than did candidate 306 due to the difficulty his/her examination form. Candidate 411 was rated by lenient judges (mean severity = -.36 logits) while candidate 306 was rated by more severe judges (mean severity = .86 logits). Similar patterns were discovered for 10 of the 74 candidates in the sample data. Cronbach's Alpha estimate could cause the results to be interpreted as "reliable", at least for the total test. However, the reality, as shown in Table 6, is that while candidates raw scores distribute normally, the assignment of those scores is linked to the judges encountered in the examination process. When judge bias is removed, the interpretation of the results may be somewhat different.

DISCUSSION

The goal of a performance examination is to make distinctions among candidate performances. All four analysis methods indicated that there is variance among topics, tasks, judges, and candidates. The Cronbach's Alpha reliability estimates show acceptable reliability for candidate total raw scores; however, those scores cannot be interpreted as independent of the judges who gave the ratings (see Table 6). The reliabilities for the other facets vary somewhat. Expectations of reliability levels for examination facets, other than candidates, have never been discussed in the literature.

The inter-judge correlations show less than perfect agreement among judges. The range of correlations was extensive due to the rotational pairing of judges among candidates. If two judges happened to have a randomly selected group of candidates in common, and both judges

Table 6
Comparison of Candidate Scores and Ability Estimates

	Candidate 411		Candidate 306	
	Raw Score	Estimated Measure(SEM)	Raw Score	Estimated Measure(SEM)
	47	.48 (.38)	47	1.70 (.38)
	Judge#	Severity	Judge#	Severity
	33	-.16	03	1.33
	04	-1.21	29	.90
	23	-.29	34	1.36
	01	.27	18	.89
	13	-.35	09	.85
	32	-.47	28	-.17
Mean Judge Severity*		-.36		.86
Mean Topic Difficulty**		.00		.00
Mean Task Difficulty**		.00		.00
Probability of Passing		70%		30%

* Lenient judges mean a higher probability of a passing score. Severe judges mean a

** All candidates challenged the same topics and tasks. Difficulty centers at 0.00

distinguished differences in candidate performance among tasks, then a correlation was calculated. However, these correlations were dramatically different ranging from -1.00 to +1.00 with an average of .00. The rotational pairing of judges interfered with using the inter-judge correlations as a means of interpreting the consistency of the judges or the candidate scores. A high correlation among judges on a subset of candidates is a questionable predictor in this examination.

The G-theory approach shows significant error variance for judges rating candidates on tasks within topics, candidates across topics, and judges rating candidates across topics. This translates to candidates challenging examinations of differing difficulty, even though the analysis does not account for those differences.

The multi-facet Rasch analysis shows variance within the facets for judges, topics, tasks, and candidates. The logit estimates for the elements within facets are shown as calibrations for judges, tasks and topics, and accounted for in the candidate ability estimates. The Reliability of Candidate Separation is .92. Candidate ability estimates are independent of the specific judges who rated specific candidates. The goal of performance examinations is to distinguish reliably among candidates' abilities, and to make the distinctions independent of the judges who rated the candidate or tasks assessed. Only the multi-facet model provides an analysis method that accomplishes this goal.

When interpreting candidate results, the analysis method used must be acknowledged, because it has a significant impact on the conclusions that may be drawn about the quality of a candidate's performance. Any method used to analyze performance data should emphasize the reliability of the interpretation of the results of the candidate's performance. The other facets of the examination are analyzed in order to better understand the candidate's results.

A major of similarity of traditional analysis methods, inter-judge correlations and G-theory, is that raw scores are used in all calculations, so that the interpretation of the results cannot be separated from the perception of the judge who rated the candidate's performance. In traditional analysis, even though Cronbach's Alpha appears to be acceptable, the raw score earned by the candidate is affected by the severity of the judge that awarded the rating. While adding up points reduces the error of measurement, it does not insure that candidates have comparable examinations.

The argument for using inter-judge reliability is that candidates can

potentially earn comparable ratings based on random assignment, if the correlation among judges is high enough. However, the inter-judge correlations are consistently less than perfect, and raw scores are used to calculate the correlation coefficients. Candidate outcomes are influenced by the severity of the judges that are randomly assigned. Also this is an inappropriate reliability calculation since it deals with the judges who are a facet of the examination, rather than the candidate performances. The rotational pattern of the judges in this examination made inter-judge correlations very difficult to calculate.

G-theory identifies sources of error variance among the facets to better estimate a candidate's true score and aid in constructing a better examination design for the **next** examination. The analysis has absolutely no impact on the outcomes or interpretations of the candidates on the current examination. The results of the study may be used to improve the design of the next study, but the current candidate outcomes are not affected. The candidates receive the raw scores they earn from the particular judges they encounter. Thus, whatever variance, due to judge bias is present, the candidate raw scores stand.

When these three methods of analysis are used, some candidates are fortunate and get lenient judges, while others are less fortunate and get more severe judges. The interpretation of candidate performance is dependent upon the characteristics of the judges encountered, whether or not this fact is acknowledged in the reporting system. The reality is that the judges are forgotten, while the interpretation of the candidate's performance stands.

The multi-facet Rasch model is the only method that accounts for the characteristics of the particular examination form encountered by a candidate. This is generally equivalent to examination equating. Interpretation of the results of the candidate's performance is independent of the particular judges that rate the performance. Table 6 shows this quite clearly. When the raw scores are used, both candidates appear to be of comparable ability; however, when the severity of the judges is accounted for, the estimated ability measures are substantially different.

The limitation of this study is that only one data set was used for the comparative analysis. However, these data are fairly representative with regard to the facets and scoring design. Similar results have been found with writing assessment (Engelhard, 1992) and clinical examination data (Lunz, Wright and Linacre, 1990). The need to make candidate ability estimates independent of particular judges is present for any data set

regardless of the facets, rating scales or candidates that are part of the process.

Judges are unique human beings, who try to be accountable to themselves and the candidates. They usually attempt to follow instructions and use the rating scale as it was intended. However, judges see the world uniquely and so have somewhat different definitions for the meaning of the categories on the rating scale and the requirement for satisfactory performance, even with training. This was shown by variable inter-judge correlations in this study and others (Burger and Burger, 1994; Lunz and Stahl, 1993; Lunz, Stahl and Wright, 1994). The Pearson correlations, G-theory, and multi-facet Rasch methods confirm error variance among the judges. Only the multi-facet Rasch method accounts for these documented facet element differences and adjusts candidate measures accordingly.

Careful planning of the examination scoring, appropriate rating scale usage, and familiarization of judges with the examination process all contribute to reliability. The precision of the candidate score or ability estimate is affected by the examination design. "Meaning" is assigned to candidate scores or ability estimates. The other examination facets contribute to the candidate's final outcome. Therefore, reliability of the candidate scores or ability estimates is the key to the defensibility of performance examinations. While the analysis methods contributed to understanding the facets of the performance examination, only the multi-facet Rasch model accounted for facet element differences in each candidate's examination before candidate ability estimates were calculated. This made the ability estimate objective and generalizable because any bias or interactions effects of the examination facets were taken into account before the ability estimate was calculated.

ACKNOWLEDGEMENT

The authors would like to personally thank Professor George Marcoulides, Department of Management Science, California State University at Fullerton, for his help and suggestions regarding the generalizability theory analysis.

REFERENCES

Blak, H. (1995). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement, 22*, 2, 41-52.

Bond, L. (1995). Unintended consequences of performance assessment: Issues of

bias and fairness. *Educational Measurement: Issues and practice, 14,* 4, 21-24.
Brennan, R.L. & Johnson, E.G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14,* 4, 9-12.
Burger, S.E. & Burger D.L. (1994). Determining the validity of performance-based assessment. *Educational Measurement Issues and Practice, 13,* 1, 9-15.
Crick, J.E. & Brennan, R.L. (1983). Manual for Genova: A generalized analysis of variance system. *The American College Testing Program, Technical Bulletin Number 43,* Iowa City, Iowa.
Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5,* 3, 171-191.
Guion, R.M. (1995). Commentary on values and standards in performance assessment. *Educational Measurement: Issues and Practice, 14,* 4, 25-27.
Koretz, D. (1992). New report on Vermont portfolio project documents challenges. *National Council on Measurement in Education Quarterly Newsletter, 1,* 4, 1-2.
LeMahiew, P.G., Gitomer, A.H. & Eresh, J.T. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice, 14,* 3, 11-28.
Linacre, J.M. (1988). *FACETS,* a computer program for analysis of examinations with multiple facets. Chicago: MESA Press.
Linacre, J.M. (1989). *Many-Facet Rasch Measurement.* Chicago:MESA Press.
Longford, N.T. (1994). Reliability of essay rating and score adjustment. *Journal of Educational and Behaviorial Statistics, 19,* 3, 171-200.
Lunz, M.E. & Stahl, J.A. (1992). New ways of thinking about reliability. *Professions Education Researcher Quarterly, 13,* 4, 16-18.
Lunz, M.E., Wright, B.D. & Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3,* 331-345.
Lunz, M.E. & Stahl, J.A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13,* 4, 425-444.
Lunz, M.E. & Stahl, J.A. (1993). Impact of examiners on candidate scores: An introduction to the use of multifacet Rasch model analysis for oral examinations. *Teaching and Learning in Medicine, 5,* 3, 174-181.
Lunz, M.E., Stahl, J.A., & Wright, B.D. (1994). Interjudge reliability and decision reproducibility. *Educational and Psychological Measurement, 54,* 4, 913-925.
Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14,* 4, 5-8.
Marcoulides. G. & Drezner, Z. (in press). *Method for analyzing performance assessment.* In Wilson, M. Draney, K., and Englehard, G. (4th Volume), *Objective Measurement Theory into Practice.* Norwood, NJ:Ablex Corp.
Rasch, G. (1960/1980). *Probability models for some intelligence and achievement tests.* Chicago: University of Chicago Press.
Wright, B. D. & Stone, M. (1979). *Best Test Design.* Chicago: MESA Press.

JOURNAL OF OUTCOME MEASUREMENT, *1*(3), 239 - 257

Interperting the Chi-Square Statistics Reported in the Many-Faceted Rasch Model

Randall E. Schumacker
University of North Texas

Mary E. Lunz
American Society of Clinical Pathologists

The different chi-square statistics reported in the many-faceted Rasch model analysis are presented and interpreted. In addition, other chi-square summary values are computed and presented for interpretation of facets. The chi-square values are useful for determining: (1) the significance of a facet in the Rasch model; (2) the significant contribution of facet main and interaction effects; (3) differences among facet elements; and (4) identifying the specific facet interaction adjustments to the subjects' calibrated logit ability measure.

Requests for reprints should be sent to Randall E. Schumacker, College of Education, Dept. of Technology & Cognition, Matthews Hall, Room 304, University of North Texas, Denton, TX 76203.

A many-faceted Rasch model analysis produces calibrated person measures that are **adjusted** for the facets included in the measurement design. Facets are measurement conditions that are hypothesized to effect a persons' score, e.g. subjects tested at different times or examinees tested over different topics or tasks. A comparison of the facets, given crossed or nested conditions, is possible even in the presence of an incomplete measurement design, i.e., a person is not rated by the same set of judges. Many-faceted Rasch models are useful for analysis of scores in many different settings including professional licensure exams and statewide testing programs where facets may affect the persons' score and no retesting under different conditions (facets) can occur, i.e., a different practical exam can't be given or students can't retake a statewide test.

From a measurement design perspective, raw scores are obtained from individuals under certain defined conditions or facets, which in the many-faceted Rasch model are converted to logit measures. Given that a particular facet element may affect subjects' scores, it is important to determine whether the facet elements are significantly different. The extent to which facet elements effect individual scores is found by examining the facet logit calibrations and noting differences among the facet elements (main effects). Additionally, one can determine if an interaction between the elements of two facets influence a subjects' score. The chi-square statistics reported by the FACETS computer program (Linacre, 1994) can be helpful in testing both these main and interaction effects of facets in many-faceted Rasch models.

Our purpose is to help the measurement specialist interpret the various chi-square values reported in the many-faceted Rasch analysis output. In addition, other helpful chi-square summary values are computed and tabled. The authors believe that these chi-square values are useful for: (1) determining the significance of a facet in the model by interpreting the overall global data-to-model fit using residuals or remaining error in full and reduced models; (2) determining the significant contribution of facet main and interaction effects; (3) determining differences among facet elements in the model; and (4) identifying the specific facet main and interaction effect adjustments to the subjects' calibrated logit measures.

METHODOLOGY

Data

A total of seventy-four (n = 74) subjects participated in the study. The facets studied included subjects, judges, sessions, topics, and tasks. The session facet was coded from 1 to 5 and represented the day of the week in which each subject was rated by a sample of six judges. The topic facet included three elements: history, geography, and earth science. The task facet included three elements: recall, interpretation, and application. A total of thirty-one (j = 31) judges rated subjects on the tasks within each topic, however, the same judges did not rate all the same subjects. There were no unplanned missing data in the measurement design.

Two different judges provided a rating of 0 = 'F', 1 = 'D', 2 = 'C', 3 = 'B', 4 = 'A' on the three tasks within each of the three topics for a given subject. Consequently, each subject received a total of eighteen (18) ratings from six (6) judges based on the tasks within the topics. Raw scores could range from 0 to 72. The sample mean = 52 with a standard deviation = 9. The FACETS computer program computed a calibrated logit ability estimate for each examinee taking into account the particular facet elements encountered by the subject. The FACETS computer program also computed separate logit estimates for the elements of the facets in the model and printed separate measurement summary tables for each facet (Appendix B).

Design

A subject was rated on three tasks in each of three topics by two judges. A subject was examined during only one of five different sessions (day of the week). The measurement design with 18 ratings indicated for only **one** subject by **six** judges is depicted in Figure 1. This measurement design is a nested design because subjects were nested within each session and not all judges rated all the same subjects. It also contained a crossed design effect with elements of the topic facet crossed with elements of the session facet. In this design, the facets analysis considers the influence of each facet upon a subject's ability estimate. From a design perspective, it was determined that the session a subject was rated in, the six judges and topics, as well as, the task difficulty, would impact the score a subject received. This resulted in a five-faceted Rasch model which included main effects for subject, session, judge, topic and task facets.

Session	1			2			3			4			5		
Topic*	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
Task**	123	123	123	123	123	123	123	123	123	123	123	123	123	123	123
Judges***															
01															
02															
03															
04															
05															
06															
07	323														
08		322													
09															
10															
11															
12															
13															
14															
15			222												
16															
17															
18															
19															
20															
21	322														
22		333													
23															
24															
25															
26															
27															
28															
29															
30															
31			233												

Note: Judges' ratings: 0='F'; 1='D'; 2='C'; 3='B'; 4='A' with subjects nested in sessions. Only 6 judges (18 ratings)for 1 subject are shown.
* Topic: a=history; b=geography; c=earth science
** Task: 1=recall; 2=interpretation; 3=application
*** Judge: consecutive numbers were assigned to each judge

FIGURE 1 Measurement design (6 judge ratings of 1 subject).

Residual Analysis

Whether a given facet contributed to the model, above and beyond the contribution of other facets in the model, can be tested using a residual chi-square which is computed by squaring and summing the standardized residual values output in a separate file by the FACETS program. The residual chi-square can indicate whether a particular facet impacted the subjects' score in the model. Basically, this can be accomplished by running separate models with only one facet missing each time. In addition, an inspection of the residual values for each subject from the full many-faceted model and the reduced many-faceted models, where only one facet was excluded, would indicate what specific influence the facet had upon each subjects' scores. Consequently, the unique effect of each facet could be examined.

Facets Analysis

The basic Rasch model (Wright & Stone, 1979) using dichotomous scoring (1=correct, 0=wrong) is depicted as:

$$\log [P_{ni1} / P_{ni0}] = B_n - D_i ,$$

where,

P_{ni1} = probability of subject n getting item i correct (x=1)
P_{ni0} = probability of subject n getting item i wrong (x=0)
B_n = ability of student n
D_i = difficulty of item i.

This basic Rasch model has two facets, subject ability and item difficulty. In many-faceted Rasch models (Linacre, 1994), this basic model is expanded to include other facets. For example, the five-faceted Rasch model used in our analysis was written as follows:

$$\log [P_{nijmsk} / P_{nijms(k-1)}] = B_n - D_i - C_j - T_m - S_s - F_k ,$$

where,

P_{nijmsk} = probability of student n being rated k on task i in topic m in session s by judge j.
$P_{nijms(k-1)}$ = probability of student n being rated k-1 on task i in topic m

in sessions by judge j.

B_n = ability of student n
D_i = difficulty of task i
C_j = effect of rating by judge j
T_m = effect of rating for topic m
S_s = effect of rating in session s
F_k = difficulty of category k, i.e., relative to category k-1

The raw scores are input and the data reformatted using a FacForm program for suitable input into the FACETS program (Linacre, 1994; see Appendix A). The FACETS program outputs several different types of chi-square values. These chi-square values are termed "fixed" effects, "random" (normal deviate), and data-to-model global "residual" fit. An understanding of each is important in making decisions about facet inclusion in a model, facet element similarity, and facet element interaction.

To examine the similarity among facet elements a "fixed" effects chi-square is computed which indicates whether the L measures (facet elements) are statistically equivalent to one common "fixed" effect apart from measurement error. The basic formula for this chi-square is (Linacre, 1994):

$$\chi^2 = \sum[(w_i D^2_i)] - (\sum w_i D_i)^2 / \sum w_i ,$$

where,

$$\text{Truevar (D)} = \sum(D_i - D_{mean})^2 / (L - 1) - [(\sum SE_i^2) / L] ,$$

$$w_i = 1 / (\text{Truevar(D)} + SE_i^2) , \text{ and degrees of freedom} = L - 2.$$

The value w_i, computed as $1/(\text{Truevar(D)} + SE_i^2)$, indicates the information for L facet element measures, D_i, with standard errors, SE_i, i.e. the information function (w_i) is multiplied times the parameter estimates (D_i) in the formula. If $p > .05$ (non-significant), then the L facet element measures are statistically equivalent indicating that all the elements are assumed equal. For example, this chi-square for testing the similarity of judges would reflect a test of the following null hypothesis (all judges are similar): H_o: $judge_i = judge_j$, where $i \neq j$.

A "random" (normal deviate) chi-square is also possible for each facet included in a model. The formula is:

$$\chi^2 = \sum (Z^2_L),$$

where,

L = the number of elements in the facet,

Z = standardized residual values.

This chi-square has an expected value equal to the number of elements in the facet. For example, topic had three elements so the expected "random" (normal deviate) chi-square value is 3. There were 31 judges so the expected "random" (normal deviate) chi-square value is 31, and so forth for the other facets.

A data-to-model global "residual" fit chi-square test is also possible where the sum of squared standardized residuals equals a chi-square value with degrees of freedom equal to the number of measurable responses minus the number of independent estimable parameters. The standardized residual values (observed minus expected score divided by the square root of the variance) is output into a separate file by the FACETS program. The number of measurable responses in the present example is 1,332 (74 subjects times 18 ratings). The data-to-model global "residual" fit χ^2 for the five faceted model is the sum of the squared standardized residuals for these 1,332 subject ratings. The data-to model global chi-square indicates overall whether the ratings fit the specific hypothesized many-faceted model. This chi-square is useful for testing the effect of including a facet in the model. For example, the chi-square with all facets in the model would represent a full model, while a reduced model excluding only one facet could be run and the chi-square values compared. Basically, if a facet doesn't have an "effect" in the model, then there should be little difference in the chi-square values, hence the standardized residual values. In all of the chi-square applications discussed, the chi-square can be converted to a linear measure for comparative purposes by taking the log (χ^2 / df) (B.D.Wright, personal communication).

RESULTS

Facet Main Effects

The five-faceted Rasch model included subjects, topics, tasks, judges, and session effects. These facets were selected based on how subjects' scores were obtained, i.e., measurement conditions. Obviously, the use of a proper measurement design is instrumental to interpreting and

understanding score results (Lunz, 1994). The "fixed" chi-square value for each facet is summarized in Table 1. These chi-square values are individually reported at the bottom of each facet measurement summary table output by the FACETS program and are reported in Appendix B with the exception of the subjects' measurement summary table. The "fixed" chi-square values were significant for all facets included in the model. This indicates that the elements for each facet, differed significantly. The interpretation of differences among facet elements relates to how the subjects' scores are affected by the particular combination of facet elements encountered by a subject, e.g., Do subjects' scores differ depending when (day of week) they were rated?. Subject ability estimates are adjusted according to the logit value assigned to the facet elements. Significant differences in facet elements indicates the need for this adjustment. The facet element logit measures for judges, sessions, topics, and tasks are presented in Appendix B.

Residual Analysis

Table 2 shows the "residual" chi-square values for the five-faceted model and the four-faceted models in which a different facet was dropped. The five-faceted model was significant at the .05 level of significance indicating that unaccounted residual variation in subject scores was still present. However, in the four-faceted reduced models, only dropping the judge facet resulted in a non-significant model at the .05 level of significance. Without the judge facet, the residual variation is reduced such that the model is no longer statistically significant which means the judge facet should not be included because it increases the magnitude of residual variation. The judge facet therefore impacted the model and affected subjects' scores making the measures less valid. An inspection of the residual values output in a residual file by the FACETS program for each subject from the five- and four-faceted models revealed the influence different judges had upon the subjects' scores. For example in Table 3, subject number two had eighteen ratings with a total observed score of 53; each observed score, expected score, residual error, variance, and standardized residual or z value are reported. The subject's raw score of 53 was converted into a 2.42 logit ability estimate with a standard error of .41 in the five-faceted model, and a 1.55 logit ability estimate with a standard error of .39 in the four faceted model, which didn't include the judge facet. The difference in the calibrated logit ability measures is due to the effect

Table 1
FACET's Main Effects

Facet	Fixed χ^2	df	p
Subjects	877.30	73	<.01
Session	12.80	4	.01
Topic	8.60	2	.01
Judge	223.90	30	<.01
Task	26.80	2	<.01

Note. Data-to-model global fit residual chi-square = 1345, df = 1217, $p < .01$.

Table 2
Full and Reduced FACET Models

Facet	Residual χ^2	df	p
Full model	1345.0	1217	.01
Reduced models:			
No Session	1312.1	1221	.03
No Judge	1315.2	1247	.09
No Topic	1320.5	1219	.02
No Task	1321.4	1219	.02

Note. Degrees of freedom is based upon the number of responses (1332) minus the number of estimable parameters.

Table 3
Residual Scores for Subject 2 in Session 1

Observed score	Expected score	Residual	Variance	Z	Topic	Judge	Task
Five Faceted Model							
4	3.27	.73	.31	1.31	1	33	1
3	3.09	-.09	.31	-.17	1	33	2
3	3.15	-.15	.31	-.27	1	33	3
3	3.14	-.14	.31	-.26	1	12	1
3	2.97	.03	.32	.06	1	12	2
3	3.03	-.03	.32	-.05	1	12	3
3	3.05	-.05	.31	-.10	2	01	1
3	2.87	.13	.33	.22	2	01	2
3	2.93	.07	.32	.11	2	01	3
3	2.87	.13	.33	22	2	29	1
3	2.67	.33	.37	.53	2	29	2
3	2.74	.26	.36	.43	2	29	3
2	2.81	-.81	.34	-1.39	3	34	1
2	2.61	-.61	.39	-.98	3	34	2
3	2.68	-.32	.37	.52	3	34	3
3	3.07	-.07	.31	-.12	3	16	1
3	2.89	.11	.33	.19	3	16	2
3	2.95	.05	.32	.09	3	16	3
Four Faceted Model (Judge excluded)							
4	3.07	.93	.35	1.58	1		1
3	2.89	.11	.37	.18	1		2
3	2.95	.05	.36	.08	1		3
3	3.07	-.07	.35	-.12	1		1
3	2.89	.11	.37	.18	1		2
3	2.95	.05	.36	.08	1		3
3	2.99	.01	.35	.02	2		1
3	2.80	.20	.39	.32	2		2
3	2.86	.14	.37	.22	2		3
3	2.99	.01	.35	.02	2		1
3	2.80	.20	.39	.32	2		2
3	2.86	.14	.37	.22	2		3
2	3.07	-1.07	.35	-1.82	3		1
2	2.89	-.89	.37	-1.47	3		2
3	2.95	.05	.36	.08	3		3
3	3.07	-.07	.35	-.12	3		1
3	2.89	.11	.37	.18	3		2
3	2.95	.05	.36	.08	3		3

Note. Z = observed score minus expected score divided by the square root of the variance. Subject ability estimate = 2.42 logits in five-facted model and 1.55 logits in four-facated model which excluded the judge facet.

judges had when rating the subjects, i.e. residual variation (error) is reduced when the judge facet is removed. Further inspection of the judge facet effect is therefore warranted to determine which judges introduced this increased residual variation when calibrating person measures.

Facet Interaction

In addition to testing for main effects of facet inclusion or differences between elements of a facet, interaction effects between the elements of two facets can be investigated. In the present example, an interaction between elements of the session (day of week) and judge facets were hypothesized due to the measurement conditions and differences in the judges' ratings. This was deemed important because six different judges rated subjects which were nested in different sessions (day of week). Consequently, the judges' ratings could differ by session thereby affecting subjects' scores. An examination of interaction is possible in the FACETS program using the individual logit estimates and associated z-scores (expected score minus observed score divided by standard error) for each combination of judge and session element. If the z-score associated with a logit measure is greater than ± 2, a significant difference exists between the "observed" score and "expected" score. A significant difference between these two scores indicates that the judge rated subjects in that session significantly different than expected based on his/her rating performance in other sessions. In Table 4, Judge 1 in the fifth session, Judge 16 in the third session, and Judge 31 in the third session indicated z-score values greater than ± 2.

Table 4 is in an **abbreviated format** and only presents the interaction logit measures for **selected** judges' ratings of **two** subjects across the session facet. To illustrate the usefulness of these results, subject number two had a raw score of 53, an estimated ability of 2.42 logits, and was rated by judges 1, 12, 16, 29, 33, and 34 in session one. The interaction effect is determined by adding the logit measures (bias measures) for these judges from session one which yields .86, i.e., (.47 + .94 + -.15 + -.67 + -.16 + .43). For comparison purposes, subject number fifteen had a raw score of 53, an ability estimate of 1.46 logits, and was rated by judges 8,12,14,26,31,and 33 in session three. The interaction effect upon this subject's score is determined by adding the logit measures for these judges from session three which yields -.20, i.e., (-.24 + -1.00 + -.77 + -.29 + 1.25 + .85). This indicates, that although these two subjects had identical raw scores, each

Table 4
Interaction Effect of Judge by Session (Subjects 2 and 15 only)

Obsvd Score	Exp. Score	Obsvd Count	Obs-Exp Average	Bias Measure	Model S.E.	Z-Score	Session	Logit Estimate	Judge	Logit Estimate
17	17.9	6	-.15	.47	.71	.67	First	.10	01	.27
21	22.7	9	-.19	.44	.49	.89	Second	.03	01	.27
22	22.2	9	-.03	.06	.50	.11	Third	-.08	01	.27
26	26.6	9	-.06	.18	.56	.32	Fourth	.22	01	.27
33	29.5	9	.38	-1.43	.71	-2.01	Fifth	-.27	01	.27
23	23.8	9	-.09	.21	.51	.41	First	.10	08	.59
20	19.9	9	.01	-.03	.46	-.06	Second	.03	08	.59
27	26.3	9	.08	-.24	.59	-.41	Third	-.08	08	.59
21	23.9	9	-.32	.69	.46	1.48	Fourth	.22	08	.59
27	24.2	9	.32	-.88	.57	-1.53	Fifth	-.27	08	.59
23	26.0	9	-.34	.94	.53	1.77	First	.10	12	.27
27	28.4	9	-.15	.49	.59	.84	Second	.03	12	.27
30	27.2	9	.31	-1.00	.61	-1.64	Third	-.08	12	.27
21	22.2	9	-.13	.30	.49	.61	Fourth	.22	12	.27
29	26.2	9	.32	-.95	.60	-1.59	Fifth	-.27	12	.27
29	27.6	9	.15	-.49	.61	-.81	First	.10	14	-1.20
25	26.9	9	-.22	.61	.54	1.14	Second	.03	14	-1.20
33	31.2	9	.20	-.77	.71	-1.10	Third	-.08	14	-1.20
24	25.1	9	-.12	.34	.55	.62	Fourth	.22	14	-1.20
31	31.0	9	.00	-.00	.62	-.00	Fifth	-.27	14	-1.20
18	17.7	6	.05	-.15	.73	-.20	First	.10	16	.51
24	21.9	9	.23	-.57	.55	-1.05	Second	.03	16	.51
21	25.3	9	-.48	1.17	.49	2.40	Third	-.08	16	.51
24	23.1	9	.10	-.25	.53	-.47	Fourth	.22	16	.51
30	28.9	9	.12	-.41	.61	-.67	Fifth	-.27	16	.51
26	27.1	9	-.13	.37	.56	.66	First	.10	26	-.44
29	30.3	9	-.15	.49	.60	.82	Second	.03	26	-.44
30	29.2	9	.09	-.29	.61	-.47	Third	-.08	26	-.44
27	26.3	9	.07	-.22	.58	-.38	Fourth	.22	26	-.44
30	28.9	9	.12	-.39	.61	-.64	Fifth	-.27	26	-.44
18	16.7	6	.22	-.67	.73	-.92	First	.10	29	.90
22	21.7	9	.04	-.08	.50	-.17	Second	.03	29	.90
25	28.1	9	-.34	.98	.55	1.79	Third	-.08	29	.90
20	19.6	9	.05	-.10	.48	-.21	Fourth	.22	29	.90
24	23.1	9	.10	-.26	.54	-.49	Fifth	-.27	29	.90
29	27.7	9	.15	-.49	.62	-.79	First	.10	31	-.58
27	27.9	9	-.10	.30	.58	.52	Second	.03	31	-.58
26	29.6	9	-.40	1.25	.57	2.17	Third	-.08	31	-.58
28	25.1	9	.32	-.97	.59	-1.62	Fourth	.22	31	-.58
29	28.7	9	.03	-.09	.60	-.15	Fifth	-.27	31	-.58
19	18.7	6	.05	-.16	.74	-.22	First	.10	33	-.16
25	24.3	9	.07	-.21	.56	-.37	Second	.03	33	-.16
26	28.4	9	-.27	.85	.58	1.46	Third	-.08	33	-.16
31	28.0	9	.34	-1.33	.70	-1.89	Fourth	.22	33	-.16
26	27.5	9	-.17	.51	.57	.90	Fifth	-.27	33	-.16
15	16.0	6	-.17	.43	.64	.68	First	.10	34	1.36
22	25.2	9	-.36	.94	.51	1.83	Second	.03	34	1.36
28	26.6	9	.16	-.46	.58	-.80	Third	-.08	34	1.36
22	20.6	9	.15	-.30	.48	-.63	Fourth	.22	34	1.36
29	27.6	9	.16	-.51	.60	-.84	Fifth	-.27	34	1.36

Note. The 31 judges were not numbered consecutively. Bias measure referes to calibrated logit measure, and only those judges rating subjects 2 and 15 are reported.

was affected differently by the judges and sessions with which they interacted, hence the different calibrated ability estimates.

The global "fixed" chi-square value reported for each measurement summary table in the FACET program can also be partitioned into the elements of each facet, e.g., for judge and/or session. The partitioned chi-square values will sum to the global (total) "fixed" chi-square value. Moreover, differences in these chi-square values can yield "simple" effects tests between levels of a facet. Table 5 presents the "fixed" chi-square values for each of the five sessions. The global fixed summative chi-square value reported indicates that the judges' ratings differed across the sessions (days of week). Four of the five individual session chi-square values were significantly different from the tabled chi-square value of 43.77 at the .05 level of significance with degrees of freedom = 30. This indicates that the judges differed significantly in their ratings within that session. Table 6 presents the "fixed" chi-square values for the thirty-one (31) judges. Judges 04, 06, 10, 13, 20, and 28 had more variation in their ratings of subjects overall as noted by the chi-square values which were significantly higher than the expected value of 5. These six judges were significantly different across the sessions and logically explain the significant interaction effect between judges and sessions. A wide variation in severe or lenient ratings by a judge across sessions will increase the chi-square value. This wide variation in judge ratings can be quickly noted by the range of Z min to Z max values. For example, judge 10 rated the sessions as follows: (1) 2.50; (2) -.32; (3) -1.56; (4) -3.00; and (5) 2.52. Judge 10 was therefore rating severely in sessions one and five, but rating leniently in sessions two, three and four. Overall, this wide variation indicates an inconsistent judge.

As noted earlier, this χ^2 value has an expected value equal to the number of facet elements, i.e., 5 because there are 5 sessions (days of the week). If no difference exists between the expected and observed scores obtained from a judges' ratings, then $\chi^2 = 0$. If χ^2 values are between 0 and 5, then a judge has given lenient ratings, i.e., observed scores are greater than expected scores. If $\chi^2 > 5$, then expected scores are higher than observed scores indicating more severe ratings. The range of z-scores, however, must be taken into consideration, and therefore minimum and maximum values are reported in Tables 5 and 6. For example, in Table 6, judge 26 is a relatively lenient rater with only small differences between expected and observed scores across the sessions, as indicated by the narrow range of z-score values (-.47 to .82). Judge 26 is consistent in rating

Table 5
Chi-Square Values for Session Facet

Session	χ^2	n of judges	Z_{min}	Z_{max}
Monday	55.42	30	.02	2.84
Tuesday	40.11	30	.00	2.60
Wednesday	61.67	31	.10	4.50
Thursday	53.12	30	.02	3.34
Friday	46.68	30	.00	2.52

Note. Global "fixed" $\sum\chi^2 = 257$, df = 150, $p < .001$
(Tabled chi-square = 43.77, df=30, p =.05 level of significance).

Table 6
Chi-Square Values for Judge Facet

Judge	χ^2	n of sessions	Z_{min}	Z_{max}
01	5.40	5	.11	-2.01
03	5.92	4	-1.31	1.88
04	29.41	4	-2.45	4.50
06	20.09	5	-2.49	3.34
07	9.25	5	-.95	2.60
08	4.87	5	-1.53	1.48
09	3.55	5	-1.06	1.05
10	24.17	5	-3.00	2.52
11	8.62	5	-1.30	2.40
12	9.43	5	-1.64	1.77
13	20.86	5	-2.44	2.84
14	3.17	5	-1.10	1.14
15	2.81	5	-1.00	1.17
16	7.79	5	-1.05	2.40
18	1.96	5	-1.14	.68
19	3.76	3	-1.34	1.40
20	14.45	5	-2.79	1.79
21	5.62	5	-1.65	1.14
22	5.84	5	-1.38	1.46
23	3.51	5	-1.26	.70
24	3.60	5	-1.19	1.33
25	4.85	5	-1.20	1.22
26	1.88	5	-.47	.82
27	7.31	5	-1.07	1.93
28	12.65	5	-2.49	1.93
29	4.36	5	-.92	1.79
30	6.35	5	-.90	2.00
31	8.25	5	-1.62	2.17
32	5.02	5	-1.28	1.22
33	6.70	5	-1.89	1.46
34	5.55	5	-.84	1.83

Note. Tabled chi-square values for df:
Chi-square = 11.07, df = 5, p = .05 level.
Chi-square = 9.49, df = 4, p = .05 level.
Chi-square = 7.82, df = 3, p = .05 level.

subjects across sessions. Judge 04, in contrast, shows large differences between expected and observed ratings across sessions (-2.45 to 4.50). Judge 4 is demonstrating a wide variation in ratings across sessions, and therefore, is less consistent.

CONCLUSIONS

The five-faceted model was hypothesized based upon a measurement design that required subject, session, judge, topic, and task facets. The main effects for each facet were examined for significance using a "fixed" χ^2 value. All facets had a significant "fixed" χ^2 value indicating that the elements of each facet were significantly different. These differences were noted when comparing the facet element logit values reported in the measurement summary tables in Appendix B. These facet element differences indicate that the elements of the facets have different effects upon the subjects' scores and need to be accounted for through an adjustment to their calibrated ability estimates.

An examination of reduced four-faceted models revealed that the judge facet impacted subjects' scores. When the judge facet was removed, the residual variation in the model was reduced. Moreover, it was noted that an examination of the range of z-score values for the judges indicated which judges had more variation or inconsistency in their ratings of subjects. Judges with larger differences between expected and observed scores in a session yielded larger z-score values. Six judges were found to have significant variation or inconsistency in their ratings of subjects. A two-way interaction between the session elements and the judges was also investigated given that judges rated subjects in different sessions (days of week). This permitted an examination of logit measures specific to the combined effect of judge and session upon a subjects' ability estimate. Depending upon which session a subject was rated in, the logit measures for the judges who rated the subject can be summed to yield the effect or adjustment that was made to the subject's ability estimate. The judges were **not** always comparable across the sessions. Six of the 31 judges showed different patterns of rating across the sessions. In future measurement designs, these six judges would require further training or possibly not be included.

We have discussed a chi-square test of facet main effects, a chi-square test that indicates if facet elements are different, a method for examining the contribution of a facet to the model, a method for examining interaction

(z-scores), and a partitioning of the global chi-square value into a "simple" effects chi-square test. These chi-square tests were presented and interpreted in the context of a measurement design. Our findings indicated that if the elements of a facet are significantly different, then the facet elements encountered by a subject should be accounted for when computing a subject's ability estimate. After all, the primary intent of the many-faceted Rasch model is not to maximize the global data-to-model fit, rather to construct generalizable linear measures for subjects taking into consideration standard error (reliability) and fit (validity).

From a measurement design perspective, the intent is to reduce measurement error and more accurately estimate subject ability (Lunz, 1994). In our example, the judge facet increased residual variation (measurement error). Being able to test whether a facet has a significant effect upon subjects' scores permits attention to properly adjusting scores. An examination of the calibrated estimates for each element of a facet (see element measures for each facet in Appendix B) indicates the particular amount of adjustment to be made to the subjects' ability estimates. When a chi-square test indicates that the facet elements differ, the facet calibrated estimates , e.g. logit estimates (bias measures in Table 4), indicate how much the subject ability estimates should be adjusted to account for the characteristics of the particular elements encountered by a subject. Herein lies the specific adjustment to a subjects' score that we seek.

ACKNOWLEDGEMENTS

Earlier version of this paper was presented in a Rasch Measurement SIG Symposium "Many-Faceted Analysis: Applications, Issues, and Techniques" on Thursday, April 11, 1996 at the American Educational Research Association annual meeting in New York, NY.

REFERENCES

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

Lunz, M.E. (1994, October). *Reducing the error of measurement by design for performance examinations.* Paper presented at the annual meeting of the Mid-Western Educational Research Association. Chicago, Illinois.

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

APPENDIX A

The original coded data was entered as follows:

ID	Session	Topic	Judge1	T11	T21	T31	Judge2	T21	T22	T23
1	1	1	1	3	4	3	2	3	4	3
1	1	2	3	4	3	2	4	2	3	3
1	1	3	5	4	4	4	6	3	3	3
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
74	5	1	11	3	3	3	9	4	3	2
74	5	2	8	2	2	3	1	3	4	4
74	5	3	31	2	2	2	10	3	3	3

The first three lines of coded data indicates one subject (ID variable) who has been rated during session one (Monday) in three topic areas. For each topic area, two different judges have provided ratings on the three tasks for a total of eighteen ratings. The last three lines indicate the last subject was rated during session five (Friday) in three topic areas by six different judges.

The Rasch Facform program converts this data set into six lines per subject with comma separated variables. The raw data for the first subject would be recoded as follows:

```
1,1,1,1,1-3,3,4,3
1,1,1,2,1-3,3,4,3
1,1,2,3,1-3,4,3,2
1,1,2,4,1-3,2,3,3
1,1,3,5,1-3,4,4,4
1,1,3,6,1-3,3,3,3
```

The values between each comma, respectively, are: subject, session, topic, judge, number of task facet elements, i.e., 1 to 3, and rating values for tasks one, two, and three. The total number of data lines in the Rasch Facform data file is n = 444 (74 subjects x 6 lines to record the 1,332 ratings) by the 31 judges.

APPENDIX B

Session Measurement Report

Session	n of ratings	Measure	S.E.	Subject Ids
Monday	252	.10	.11	1-14
Tuesday	270	.03	.10	15-29
Wednesday	270	-.08	.10	30-44
Thursday	270	.22	.10	45-59
Friday	270	-.27	.11	60-74
Total	1332	.00		

Note. Fixed $\chi^2 = 12.8$, df=4, p = .01

Task Measurement Report

Task	n of ratings	Measure	S.E.
Recall	444	-.32	.08
Interpretation	444	.25	.08
Application	444	.07	.08
Total	1332	.00	

Note. Fixed $\chi^2 = 26.8$, df=2, p<.01

Topic Measurement Report

Topic	n of ratings	Measure	S.E.
History	444	-.09	.08
Geography	444	.19	.08
Earth Science	444	-.10	.08
Total	1332	.00	

Note. Fixed $\chi^2 = 8.6$, df=2, p=.01

Judge Measurement Report

Judge	n of ratings	Measure	S.E.
01	42	.27	.26
03	33	1.33	.26
04	30	-1.21	.39
06	48	-.61	.25
07	45	-.50	.26
08	45	.59	.23
09	45	.85	.24
10	45	.24	.25
11	45	.59	.23
12	45	.27	.25
13	45	-.35	.26
14	45	-1.20	.27
15	45	.25	.24
16	42	.51	.25
18	45	.89	.22
19	27	-.94	.34
20	45	.00	.25
21	45	-.23	.25
22	42	-.90	.29
23	45	-.29	.26
24	45	-.81	.26
25	48	.48	.22
26	45	-.44	.27
27	45	.90	.23
28	42	-.17	.26
29	42	.90	.24
30	42	-.56	.27
31	45	-.58	.26
32	45	-.47	.26
33	42	-.16	.27
34	42	1.36	.25
Total	1332	.00	

Note. Fixed χ^2 =223.9, df=30, $p < .01$

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.

 ISSN 1090-655X.