Volume 1, Number 1, 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 1 1997

Editor's Note

Journal of Outcome Measurement is intended to provide a multi-disciplinary perspective on the theoretical and applied aspects of objective measurement. Outcomes, in the sense intended in the title, are the results of any planned intervention and should be interpreted in its broadest sense. The journal will strive for approximately equal representation of work in education/psychology and health sciences, with a primary focus on the ground breaking work in physical medicine and rehabilitation. However, articles relating to other areas will be considered if there are methodological implications for the primary focus of the journal. As many in the social sciences turn towards outcomes-based measures of productivity and payment the availability of resources such as this journal will be extremely critical. As other fields try to duplicate the successes in physical medicine and rehabilitation, there needs to be a chronicle of the work being done in this field. The journal will also strive for a balance between applied and theoretical articles.

This journal is the only journal devoted specifically to the theory and applications of objective measurement. As such, it will be a logical source for new practitioners trying to adopt the outcomes work in physical medicine and rehabilitation to new fields. Although there are other measurement journals, Journal of Educational Measurement, Applied Educational Measurement, Applied Psychological Measurement, Educational and Psychological Measurement, and Psychometrika, these journals cover many competing measurement models and have a primary focus in psychology and education. Medical journals, such as the Archives of Physical Medicine and Rehabilitation, The Journal of Rehabilitation, and the American Journal of Physical Medicine and Rehabilitation also publish articles on topics that would be covered by the Journal of Objective Measurement, these journals remain the primary source of research in the medical speciality. Because measurement articles on outcomes currently compete for space in medical journals, the amount of information available for both clinical and measurement issues is limited.

Richard M. Smith
Rehabilitation Foundation, Inc.

JOURNAL OF OUTCOME MEASUREMENT, *1*(1), 2-18

Establishing the Diagnostic Validity of Premenstrual Dysphoric Disorder Using Rasch Analysis

Sarah Gehlert
The University of Chicago

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

Premenstrual Dysphoric Disorder (PMDD) has remained in appendices of the last two editions of *The Diagnostic and Statistical Manual of Mental Disorders* due to lack of empirical study. Items included in its set of research criteria are considered tentative pending evidence of diagnostic validity. The present study attempts to establish the construct validity of the PMDD criteria using the Rasch method to analyze the validity of individual items as contributors to the diagnosis, in contrast to the usual but less precise approach of using an external validator to establish the diagnostic utility of psychiatric conditions. Analysis of which items best differentiate participants with and without PMDD provides an idea of the relative ability of these items to distinguish PMDD. It is recommended that the areas of anger/irritability, depressed mood, and problems in interpersonal functioning be expanded in further studies and corresponding items added to symptom checklists.

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

That premenstrual symptoms affect women's mood and behavior has been acknowledged since Hippocrates (Simon, 1978). In 1931, Frank used the term "premenstrual tension" to describe the feelings that some women experience from 10 to 7 days preceding menstruation until the menstrual flow begins (Frank, 1931). In the early 1950s, Greene and Dalton (1953) coined the term "premenstrual syndrome" (PMS) to describe a condition in which women had more premenstrual symptoms than tension alone. Since that time, PMS has received much popular attention. Research on the condition has, however, been hampered by definitional ambiguity. Over 150 symptoms have been attributed to the condition but no agreed-upon definition exists (Rubinow & Roy-Byrne, 1984).

In the late 1970s and early 1980s, the idea arose that there was a subtype of PMS that was characterized primarily by severe debilitating mood disturbances. The new disturbance was thought to occur in far fewer women than did PMS. While both PMS and PMDD included physical and affective symptoms, in PMDD the latter were thought to predominate. In an effort to avoid the definitional ambiguities of PMS, the American Psychiatric Association formulated a definition of the new condition, then named Late Luteal Phase Dysphoric Disorder (LLPDD), and included it in an appendix of *DSM-III-R* (revised third edition of the *Diagnostic and Statistical Manual of Mental Disorders,* American Psychiatric Association, 1987). By mid-1994, when *DSM-IV* (American Psychiatric Association, 1994) was published, the criteria for LLPDD had been altered and it had been renamed Premenstrual Dysphoric Disorder (PMDD). It remained in an appendix of *DSM-IV*.

PMDD is defined by four research criteria. The first criterion includes at least 5 of 11 possible symptoms, one of which is affective. Symptoms must be present for most of last week of the luteal phase and be absent in the week postmenses. The second criterion is that PMDD must interfere markedly with school, work, or interpersonal relationships and the third that its symptoms cannot represent an exacerbation of another psychiatric disorder. The fourth criterion requires that the first three criteria must be confirmed by prospective daily ratings for two consecutive menstrual cycles.

The move from the appendices to the body of *DSM* hinges on the accrual of data establishing the diagnostic validity of PMDD (Gold, 1994). In *DSM-IV,* proposed diagnostic categories are considered sets of research criteria in need of refinement. The items contained in these sets are viewed as tentative, and, as such, subject to alteration. Items in the present criteria

set could be reconfigured or dropped, or new items could be added.

Establishing the diagnostic validity of criteria sets is essential to the development of a psychiatric nosology. Methods of classifying psychiatric disorders have changed through time (Kendler, 1990). In the previous century, preeminent men in psychiatry such as Kraepelin (1893) developed and disseminated their own diagnostic systems. From the turn of the century to the 1980s, psychiatric disorders were defined by national or international organizations based on the consensus of experts. With the development of *DSM-III* (American Psychiatric Association, 1980), the emphasis shifted to basing decisions about psychiatric nosology on empirical method.

Robins and Guze's approach (Feighner et al., 1972; Robins & Guze, 1970) has become the standard for establishing the validity of psychiatric diagnoses. Here, external validators are used to establish diagnostic validity based on their ability to differentiate groups. Patients with diagnoses of schizophrenia, for example, might be divided into groups based on predicted outcomes of their conditions and the predictions validated in follow-up studies.

Kendler (1990) lists the potential for disagreement among validators as a major problem with using external validators. He notes that this and other scientific methods in use leave the researcher with no clear direction regarding which criteria set most accurately describes a certain psychiatric disorder.

Recent work on psychological assessment, although it focuses on scores on psychological tests rather than proposed psychiatric diagnoses, provides some guidance through its concern with validating hypothetical constructs. Foster and Cone (1995) say that content validity and accuracy are key to establishing whether a test really measures the latent construct that it claims to measure. Content validity, a category of construct validity, is concerned with whether elements of a test represent the construct being measured for a particular purpose. This is especially difficult to determine in the case of poorly defined constructs (Murphy & Davidshofer, 1994). Accuracy has to do with the extent to which scores on the test capture the behavior in question. It is established by comparison with observation or other evidence of the behavior. Such a comparison would not be possible in the case of PMDD, for two reasons. First, its symptoms, on the whole, represent covert rather than observable behavior. Second, exactly what behaviors constitute PMDD has not been established.

Several procedures for establishing content validity have been proposed (see e.g., DeVellis, 1991), many of which seem to have been observed in

developing the research criteria for PMDD. Of concern is the adequacy of external validation, or criterion-related validity, to further determine whether items in *DSM-IV* research criteria are relevant to the PMDD construct. As Haynes, Richard, and Kubany (1995, p. 245) have noted, high magnitudes of shared variance between scores on newly developed instruments and criteria may result from variance in items outside the domain of the new construct. It is also possible for the criterion instrument to contain items outside of the domain of the new construct, which would depress criterion-validity scores.

The present study is an approach to establishing the construct validity of the *DSM-IV* criteria set of PMDD by analyzing the validity of individual items as contributors to the diagnosis rather than by using an external validator. The Rasch method (Rasch, 1980; Wright & Masters, 1982; Wright & Stone, 1979), which is more commonly used for scale construction, is used for this analysis. The approach holds promise for providing direction in an area of psychiatry in which, at present, little direction exists, namely determining whether diagnostic criteria proposed by experts represent the conditions they mean to represent. The method is meant to be used in conjunction with approaches like Robins and Guze's (1970) and their elaborations (Kendler, 1980).

METHODS

Participants

Participants were 117 women of reproductive age who were neither pregnant, naturally menopausal (i.e., no menstrual period for one year), nor had had an oophorectomy. They were recruited for a study of changes in women's health through time from outpatient clinics of an urban teaching hospital. Obstetrics and gynecology clinics purposefully were excluded in order to de-emphasize the menstrual-cycle focus of the study. Women between the ages of 13 and 55 years who checked in for clinic appointments were given letters soliciting their participation. Fourteen of the 117 women who were recruited dropped out prior to completion of the study and four participants provided unusable data, leaving 99.

Participants were assessed for PMDD using the research criteria set from *DSM-IV* by methods outlined in an earlier study by the authors (Gehlert, Hartlage, & Chang, in press). Nine of 99 study participants met the diagnosis of PMDD using *DSM-IV* research criteria.

Instruments

After a period of instruction and orientation to the study, each participant responded to a daily symptom and mood checklist containing 24 items or subsymptoms derived from the 11 symptoms listed in *DSM-IV*. The items were rated on a 6-point rating scales (from 0 = "I did not experience the symptom at all" to 5 = "I experienced the symptom very strongly") with the intervening points indicating the increasing intensity with which the participant experienced the symptom. Daily rating data from seven post-menses follicular days (the week postmenses) and the seven days of the late luteal phase (the week before menses) were analyzed. Women were instructed to fill out the daily symptom and mood checklist at the same time each morning. They were given $80.00 upon completing the study.

Analysis

The psychometric technique was rating scale analysis (Wright & Masters, 1982), an extension of the family of measurement models devised by Danish mathematician Georg Rasch (Rasch, 1980; Wright & Stone, 1979). The rating scale model specifies that the log odds of scoring in the greater of two adjacent categories is a function of three additive parameters: person measure, item difficulty, and step difficulty. The log odds is given by:

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

in which P_{nij} is the probability of person n scoring in category j of item i, $P_{ni(j-1)}$ is the probability of person n scoring in category j-1 of item i, B_n is the measure of person n, and D_i is the difficulty of item i, and F_j is the difficulty of the step from category j-1 to category j. In the present study, F_1 is the transition from category 0 to category 1 and F_5 is the transition from category 4 to category 5. The BIGSTEPS computer program (Wright & Linacre, 1995) was used for Rasch analyses. Separate item calibrations were obtained for the non-PMDD and PMDD groups. Daily ratings data from the follicular and late luteal phases were calibrated for each group. The non-PMDD group produced data from 90 persons for 14 days, or 1,260 records. The PMDD group produced data from nine persons for 14 days, or 126 records. Differential item functioning (DIF) detecting procedures (Wright & Stone, 1979; Smith, 1996) were applied to the item difficulties obtained in

order to determine whether the items were experienced by the two groups in the same way for the two menstrual cycles. Four identity plots with 95% confidence intervals were drawn for pairs of calibrated item difficulties, comparing each phase for each group with each other phase and group. This graphical approach was used to detect item difficulties that differed between the two groups of participants and between the two phases. The approach provides a clear picture of differential item functioning. Items are modelled to have the same relative difficulty on both occasions, that is item difficulties are modelled to fit an identity line from the lower left to the upper-right hand corners of the plots. Differentially functioning items appear as outliers when they fall outside the identity line confidence bands.

Unweighted item fit mean square (MNSQ) values were also calculated in order to identify potential misfitting items. Items with unweighted fit mean square values higher than 1.2 were identified as possible misfitting items according to Rasch models (Wright & Linacre, 1995).

Items that fell outside the 95% confidence limits in the identity plots were eliminated so that further analysis could explore the relationships of items that more closely describe the PMDD construct. The remaining items were calibrated using data from the follicular phase of women in the non-PMDD group, since this phase is the most quiescent and stable of the four phases under study (Endicott et al., 1986). The item difficulty calibrations and step calibrations obtained from this group and phase were then used to anchor the measurement frame of reference. All data were then used to reestimate person measures: (a) to see if participants with PMDD would emerge as misfitting and (b) to determine the percentage of participants in each group whose person measures exceeded the item difficulty measures for each item.

RESULTS

Summaries of item difficulties by phase of cycle (follicular and late luteal) and group (non-PMDD and PMDD) are in Table 1. The hierarchies of item difficulties are different for the two groups during each of the two phases. Item 10B (a sense of being "out of control") was the least severely experienced item for both phases of the non-PMDD group. For the PMDD group, item 11 (physical symptoms) was the least severe item reported during the follicular phase and item 1C (feelings of hopelessness) was the least severe item reported during the late luteal phase.

TABLE 1

Item Difficulties and Standard Errors (SE) for Subsymptoms of Premenstrual Dysphoric Disorder (PMDD) by Phase of Cycle (Follicular and Late Luteal) and Group (Non-PMDD and PMDD)

		Non-PMDD				PMDD			
		Follicular		Late luteal		Follicular		Late luteal	
Symptom	Subsymptom(s)	D	SE	D	SE	D	SE	D	SE
One	Self-deprecating thoughts (1A)	0.06	0.05	0.09	0.04	0.00	0.10	0.06	0.10
One	Marked depressed mood (1B)	0.05	0.05	0.16	0.05	0.03	0.10	0.26	0.11
One	Feelings of hopelessness (1C)	0.41	0.06	0.34	0.05	0.37	0.11	0.58	0.12
Two	Marked anxiety (2A)	-0.36	0.04	-0.38	0.04	-0.32	0.09	-0.49	0.10
Two	Marked tension (2B)	-0.41	0.04	-0.27	0.04	-0.25	0.09	-0.34	0.10
Two	Feelings of being "on edge" (2C)	0.01	0.04	0.02	0.04	-0.10	0.09	-0.15	0.10
Two	Feelings of being "keyed up" (2D)	-0.03	0.04	0.00	0.04	-0.02	0.10	0.06	0.10
Two	Increased sensitivity to rejection (3A)	0.06	0.05	0.03	0.04	0.02	0.10	0.01	0.10
Three	Sudden sadness or tearfulness (3B)	0.02	0.04	0.12	0.04	-0.16	0.09	0.33	0.11
Four	Persistent and marked anger (4A)	0.38	0.06	0.35	0.05	0.15	0.10	0.14	0.10
Four	Increased interpersonal conflicts (4B)	0.32	0.05	0.29	0.05	0.47	0.11	0.31	0.11
Four	Persistent and marked irritability (4C)	0.28	0.05	0.09	0.04	0.13	0.10	-0.10	0.10
Five	Decreased interest in usual activities (5)	0.22	0.05	0.13	0.04	0.00	0.10	0.25	0.11
Six	Sense of difficulty in concentrating (6)	-0.08	0.04	-0.02	0.04	-0.24	0.09	-0.04	0.10
Seven	Easy fatigability (7A)	-0.44	0.04	-0.29	0.04	-0.31	0.09	-0.45	0.10
Seven	Marked lack of energy (7B)	-0.38	0.04	-0.31	0.04	-0.32	0.09	-0.20	0.10
Eight	Marked change in appetite (8A)	-0.14	0.04	-0.13	0.04	-0.38	0.09	-0.49	0.10
Eight	Overeating (8B)	-0.07	0.04	-0.21	0.04	-0.04	0.10	-0.29	0.10
Eight	Craving for specific foods (8C)	0.01	0.04	-0.06	0.04	0.17	0.10	-0.39	0.10
Nine	Hypersomnia (9A)	0.27	0.05	0.22	0.05	0.06	0.10	0.24	0.10
Nine	Insomnia (9B)	0.01	0.04	0.03	0.04	0.19	0.10	0.40	0.11
Ten	Sense of being overwhelmed (10A)	-0.28	0.04	-0.15	0.04	-0.35	0.09	-0.21	0.10
Ten	Sense of being "out of control" (10B)	0.58	0.06	0.55	0.06	0.34	0.11	0.39	0.11
Eleven	Physical symptoms (11)	-0.47	0.04	-0.58	0.03	0.54	0.12	0.15	0.10

Note. Lower item difficulty measures indicate items that were experienced more severely.

The two groups of participants experienced different item severities during their follicular and late luteal phases. In order to better understand these variations, four identity plots were drawn. Item difficulties were plotted against one other in each plot, with a pair of 95% quality control lines added to illustrate how satisfactorily the item points in the plot followed the expected identity line and to show which items departed from the identity region.

In Figure 1, item difficulties obtained from the follicular phase of the non-PMDD group were plotted against those from the group's late luteal phase. This shows whether individual items operate differently during the two phases. As can be seen, item difficulties for the non-PMDD group's two phases were comparable. They demonstrated almost the same relative difficulty. Only one of this group's 24 items operated differently during the two phases. In contrast, three items operated significantly differently during the PMDD group's phases (see Figure 2). Sudden sadness or tearfulness was experienced less severely during the late luteal phase, while two somatic items, a conglomerate of physical symptoms and craving for specific foods were experienced less severely during the follicular phase.

Item difficulty plots were also drawn to evaluate whether the non-PMDD and PMDD groups experienced different item severities during their follicular and their luteal phases. Most items fell within the 95% control lines during the follicular phases, with two exceptions (see Figure 3). Item 8A (marked change in appetite) was experienced slightly less severely by the non-PMDD group. Item 11 (physical symptoms) departed appreciably from the identity line and fell far outside the 95% confidence bands. It was the least severely experienced item for the PMDD group, but nearly the most severely experienced item for the non-PMDD group. In the item difficulty plot for the late luteal phases of the two groups, items 8A, 8C, 9B, and 11 fell outside the 95% confidence bands (see Figure 4). Items 9B (insomnia) and 11 (physical symptoms) were more difficult for the PMDD group to endorse, or less severely experienced by the group, while item 8C (craving for specific foods) and 8A (marked change in appetite) were more difficult for the non-PMDD group.

Lastly, potential misfitting items were examined by phase and group according to the rule that items with an unweighted item fit mean square value higher than 1.2 were misfitting items (see Table 2). Ten of 24 items (42%) were identified as misfitting for the follicular and late luteal phases of the non-PMDD group. Ten of 24 (42%) items for the PMDD group's follicular phase were identified as misfitting, as were 11 of 24 (46%) for the group's late luteal phase.

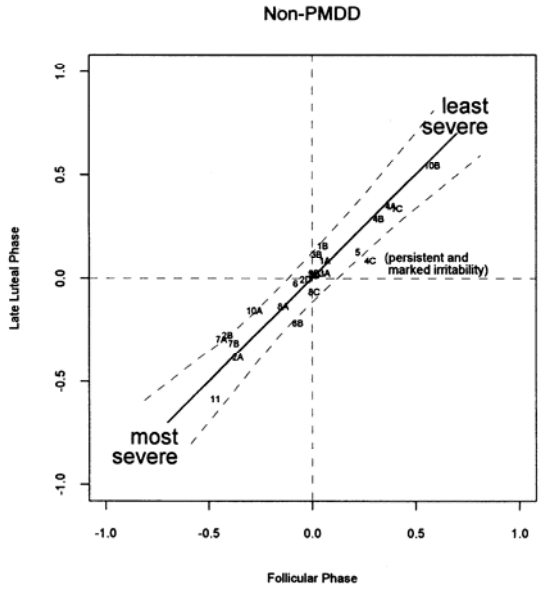


FIGURE 1 Estimates of the difficulty of the 24 items of *DSM-IV* research critieria for Premenstrual Dysphoric Disorder (PMDD) for the two phases (follicular and late luteal) of non-PMDD participants with severity of experience indicated.

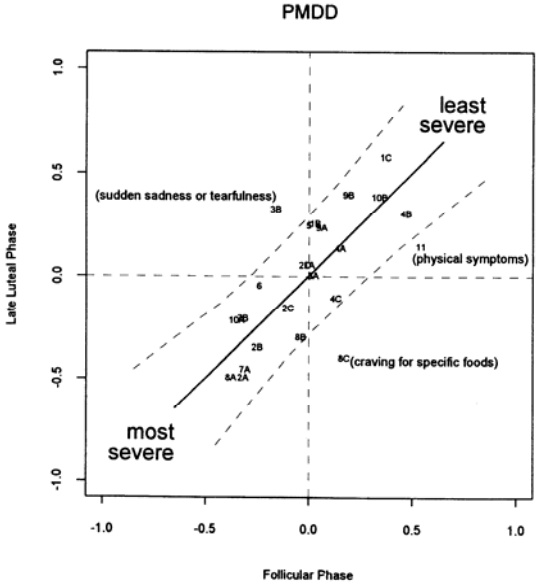


FIGURE 2 Estimates of the diffiuclty of the 24 items of *DSM-IV* research criteria for Premenstrual Dysphoric Disorder (PMDD) for the two phases (follicular and late luteal) of PMDD participants with severity of experience indicated.

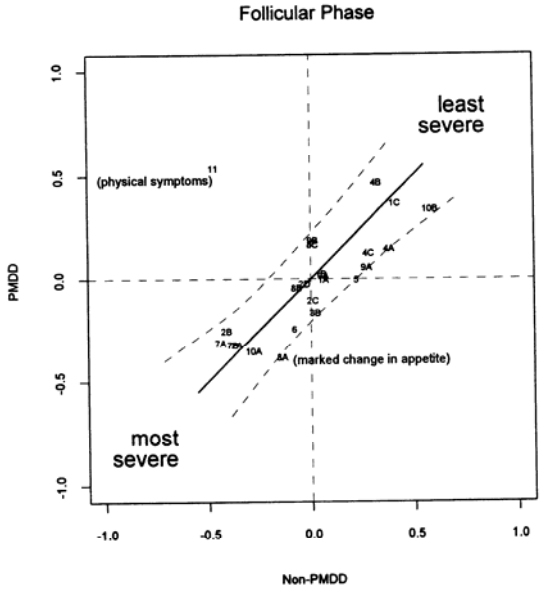


FIGURE 3 Follicular phase estimates of item difficulty for the 24 items of DSM-IV research criteria for Premenstrual Dysphoric Disorder (PMDD) by group (PMDD and non-PMDD) with severity of experience indicated.

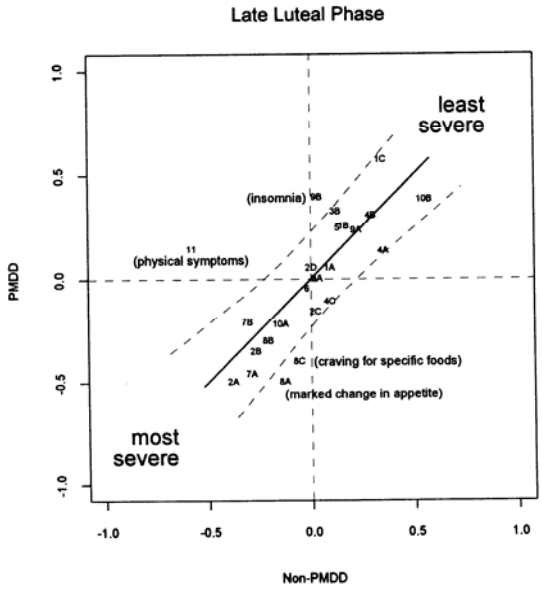


FIGURE 4 Late luteal phase estimates of item difficulty for the 24 items of *DSM-IV* research critieria fo rPremenstrual Dysphoric Disorder (PMDD) by group (PMDD adn non-PMDD) with severity of experience indicated.

TABLE 2

Unweighted Item Fit Mean Squares (MNSQ) for Subsymptoms of Premenstrual Dysphoric Disorder (PMDD) by Phase of Cycle (Follicular & Late Luteal) and Group (Non-PMDD & PMDD)

		Non-PMDD		PMDD	
Symptom	Subsymptom(s)	Follicular	Late luteal	Follicular	Late luteal
One	Self-deprecating thoughts (1A)	0.98	0.81	0.67	0.74
One	Marked depressed mood (1B)	0.65	0.82	**1.35**	0.88
One	Feelings of hopelessness (1C)	0.64	0.77	0.89	0.73
Two	Marked anxiety (2A)	1.01	**1.58**	1.10	0.99
Two	Marked tension (2B)	0.64	0.66	0.65	0.59
Two	Feelings of being "on edge" (2C)	0.66	0.53	0.60	0.69
Two	Feelings of being "keyed up" (2D)	**1.35**	**1.22**	0.81	0.80
Three	Increased sensitivity to rejection (3A)	0.71	0.81	0.91	0.77
Three	Sudden sadness or tearfulness (3B)	0.99	1.10	**1.62**	1.16
Four	Persistent and marked anger (4A)	0.81	0.99	0.63	0.64
Four	Increased interpersonal conflicts (4B)	0.83	0.75	0.92	0.55
Four	Persistent and marked irritability (4C)	0.67	0.64	0.66	0.63
Five	Decreased interest in usual activities (5)	1.10	0.94	0.78	1.02
Six	Sense of difficulty in concentrating (6)	0.76	0.85	1.11	0.79
Seven	Easy fatigability (7A)	0.91	1.03	0.91	1.19
Seven	Marked lack of energy (7B)	0.88	0.96	0.81	1.08
Eight	Marked change in appetite (8A)	**1.77**	1.08	**1.45**	**1.31**
Eight	Overeating (8B)	**1.70**	**1.72**	1.10	**2.24**
Eight	Carving for specific foods (8C)	**1.26**	1.11	**1.70**	**1.90**
Nine	Hypersomnia (9A)	**1.66**	**1.69**	**2.29**	**1.33**
Nine	Insomnia (9B)	**1.46**	**1.44**	**1.26**	**1.66**
Ten	Sense of being overwhelmed (10A)	1.13	0.74	0.95	**1.38**
Ten	Sense of being "out of control" (10B)	0.61	0.60	0.87	0.83
Eleven	Physical symptoms (11)	**1.69**	**1.71**	**2.37**	1.07

Note. Unweighted item fit mean square $\geq$ 1.20 in bold.

Items that fell outside the 95% confidence limits in the identity plots, namely items 3B, 8A, 8B, 9B, and 11, and one item with a particularly high unweighted item fit mean square value (9A), were eliminated from further analysis. Item difficulty measures of the remaining 18 items, obtained after calibration using data from the follicular phase of women in the non-PMDD group and anchoring for the whole data set, were evenly dispersed and ranged from -0.44 to 0.58 (see Table 3). One hundred and sixty of the 1,260 possible patient records (12%) were identified as misfitting, by the criteria of an absolute unweighted person fit mean square equal or greater than 2. Forty-three and four tenths percent of the PMDD group misfit, whereas only 9.1% of the non-PMDD group was identified as misfitting, The nine participants in the PMDD group had a mean of 5.9 out of 14 days of misfitting daily symptom data. One hundred and seven misfitting patient records were identified for the non-PMDD group. Table 3 displays item difficulty in ascending order as well as the percentages of person measures that exceed item difficulties. Here we see that the percentage of items for which person measure exceeded item measures was consistently higher for the PMDD group.

DISCUSSION

Interesting patterns emerge from this study that shed light on which items of the research criteria for PMDD are more and less central to diagnosis. This provides direction for a reevaluation of how the diagnostic category is configured. After discussing what the results of the present study suggest for such a reconfiguration, additional steps toward achieving a valid diagnostic category will be proposed.

An examination of item difficulties and unweighted item fit mean squares of the unreduced item set provides a gross picture of which items are more or less central to the diagnosis. Two patterns emerged among the items that operated differently between phases or groups. The first pattern is that some items were consistently misfitting items, with one exception. Sudden sadness or tearfulness misfit in all phases except for the follicular phase of the non-PMDD group. The second pattern is that all but one of these items can be described as physiological or somatic. The exception is again sudden sadness or tearfulness. Further examination of all items that could be considered physiological or somatic, those that correspond to symptoms 8, 9, and 11, shows that all were identified as misfitting in both phases of both groups. In fact, when these items are eliminated from

TABLE 3

Item Difficulty of 18 Items of Premenstrual Dysphoric Disorder and Percentage of Persons Whose Person Measure Exceeds That Item Difficulty By Group (Non-PMDD and PMDD)

Symptoms	*Item Difficulty*	*Non-PMDD (n=1260)*	*PMDD (n=126)*
Marked anxiety (2A)	-.44	7.8	53.3
Easy fatigability (7A)	-.43	7.8	53.3
Marked tension (2B)	-.40	7.4	51.6
Marked lack of energy (7B)	-.40	7.4	51.6
Sense of being overwhelmed (10A)	-.27	5.0	40.2
Overeating (8B)	-.18	4.2	40.2
Sense of difficulty in concentrating (6)	-.09	2.8	32.0
Feelings of being "keyed up" (2D)	-.02	2.1	27.9
Feelings of being "on edge" (2C)	-.02	2.1	27.9
Increased sensitivity to rejection (3A)	.03	2.0	24.6
Self-deprecating thoughts (1A)	.06	1.8	24.6
Marked depressed mood (1B)	.11	1.4	21.3
Persistent and marked irritability (4C)	.16	1.4	18.9
Decreased interest in usual activities (5)	.17	1.1	18.0
Increased interpersonal conflicts (4B)	.35	0.7	9.8
Persistent and marked anger (4A)	.36	0.7	9.8
Feelings of hopelessness (1C)	.44	0.4	4.1
Sense of being "out of control" (10B)	.58	0.3	3.3

Note. Lower item difficulty measures indicate items that were experienced more severely.

calculation of the percentage of items that misfit, a markedly different picture emerges. Only twenty-two percent of items for the two phases of the non-PMDD group and the PMDD group's follicular phase and 28% of items for the PMDD group's late luteal phase misfit. Thus, eliminating physiological symptoms decreases the percentage of misfitting items by almost half. It is also the case that the physiological or somatic symptoms were the only ones to consistently misfit across phases and groups and to yield mean square misfit statistics higher than two.

A second pattern emerges when the unweighted item fit statistics of affective or emotional symptoms are considered. Several symptoms have only one or no misfitting items. Notable is symptom 4, which has to do with anger, conflicts, and irritability, for which no items misfit. This suggests an item that fits the diagnosis well. This is also true of symptom 1, with one exception. This symptom, which has to do with depressed mood, yielded only one of 12 possible misfit statistics, which was for the PMDD group's follicular phase. The only misfit statistic for symptom 6, which has only one item and has to do with decreased ability to concentrate, was also for the PMDD group's follicular phase. These symptom might then, too, be considered more central to the diagnosis of PMDD.

It is, of course, possible that items are incorrectly configured into symptoms in the criteria set for PMDD. With this in mind, individual items in symptoms remaining after the above discussion were examined. Two items of symptom 2, namely marked tension and feelings of being "on edge," failed to misfit, as did the first of the two items of symptom 3 (increased sensitivity to rejection). The same was true for the second of the two items of symptom 10, sense of being "out of control."

A more precise picture emerges when person-fit statistics are introduced after item misfit reduction. The two groups differ appreciably in their percentages of misfitting persons, for instance. Only 9.1% of the non-PMDD group were identified as misfitting, while 43.4% of the PMDD group misfitted. Also of note is that the PMDD participants shared a mean of 5.9 misfitting days. This suggests a confirmation of their membership in the PMDD group. That some non-PMDD participants misfit reflects their meeting some, but not all, criteria for the diagnosis.

Analyses of which items best differentiated participants with and without PMDD shed additional light on which items are more central to the diagnosis of PMDD. The PMDD group experienced both easy and difficult items more severely than did the non-PMDD group (see Figure 3). Over 90% of non-PMDD participants failed to experience any of the 18 items severely. These

items were able to differentiate the two groups, and can, therefore, be considered generally good indicators of the diagnosis. Table3 provides an idea of the relative capacity of the 18 items to distinguish PMDD. All three items from symptoms 1, 4, and 5 perform the task well. These symptoms have to do with depressed mood, anger/irritability, and decreased interest. They fall immediately behind sense of being "out of control," which performs much better than does the other item of symptom 10, sense of being overwhelmed.

Support for the above picture comes from a comparison of the results of Rasch analyses with those of a factor analysis done on the same sample (Gehlert, Chang, & Hartlage, 1996). Strong similarities between the two analyses are noted. In the earlier study, tension (2B) and being "on edge" (2C), irritability (4C), anger (4A), and increased interpersonal conflict (4B) were found to be important to the definition of PMDD. This is also the case in the present study. Sleep problems (9A and 9B) were found to be less important to the definition, which was also true in the present study.

Rasch measurement seems a promising approach to the ongoing process of validating psychiatric diagnoses. A limitation of the present study, and no doubt one for all psychiatric prevalence studies, is small sample size. It should be noted, however, that this problem was somewhat lessened by the longitudinal nature of the data and Rasch's ability to consider records rather than cases. For this reason, measurement error is of less concern than selection bias, namely that these participants with PMDD might in some way differ from the larger pool of women with PMDD.

Several recommendations for altering the items of the PMDD criteria set in subsequent prevalence studies are suggested. Because somatic or physiological symptoms do not appear to be essential to the diagnosis, this area of inquiry should not be expanded. In order to ensure that the domains represented by items that are essential to the diagnosis are adequately represented, the areas of anger/irritability, depressed mood, and problems in interpersonal functioning should be expanded, and appropriate items added to symptom checklists. Further exploration of the area of interpersonal problems is indicated, as is the area of overt and covert hypomanic behavior. It is interesting to note that Kraepelin (1893, p. 353) included menstruation among the causes of a manic form of periodic mental disturbance in an early form of his nosology. This notion, which was dropped in later editions, may have had merit. Because the items of symptom 10 do not operate together, it might be advisable to reconstruct the symptom. Sense of being "out of control" fits better with symptom 4, which has to do with anger, irritability,

and conflict, or could become part of a new hypomanic constellation. Increased sensitivity to rejection (3A), which lost its partner in symptom 3 to the item reduction process, might either become part of symptom 1, which describes depressed mood, symptom 2, which has to do with anxiety and tension, or, perhaps, a new symptom involving problems in interpersonal functioning.

REFERENCES

American Psychiatric Association (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Washington, DC: Author.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

DeVellis, R. F. (1991). *Scale development: Theory and applications.* Newbury Park, CA: Sage.

Endicott, J., Nee, J., Cohen, J., & Halbreich, U. (1986). Premenstrual changes: Patterns and correlates of daily ratings. *Journal of Affective Disorders, 10,* 127-135.

Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry, 26,* 57-63.

Foster, S. L., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment, 7,* 248-260.

Frank, R. T. (1931). The hormonal basis of premenstrual tension. *Archives of Neurology and Psychiatry, 26,* 1053-1057.

Gehlert, S., Chang, C., & Hartlage, S. (1996). Symptom patterns of the research criteria for premenstrual dysphoric disorder. Manuscript submitted for publication.

Gehlert, S., Hartlage, S. & Chang, C. (in press). A design for studying the DSM-IV research criteria of premenstrual dysphoric disorder. *Journal of Psychosomatic Obstetrics and Gynecology.*

Gold, J. H. (1994). Historical perspectives of premenstrual syndrome. In J. H. Gold & S. K. Severino (Eds.), *Premenstrual dysphorias: Myths and realities* (pp. 171-183). Washington, DC: American Psychiatric Association Press, Inc.

Greene, R., & Dalton, K. (1953). The premenstrual syndrome. *British Medical Journal, 1,* 1007-1014.

Haynes, S., Richard, D., & Kubany, E. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238-247.

Kendler, K. S. (1980). The nosologic validity of paranoia (simple delusional disorder): A review. *Archives of General Psychiatry, 37,* 699-706.

Kendler, K. S. (1990). Toward a scientific psychiatric nosology: Strengths and limitations. *Archives of General Psychiatry, 47,* 969-973.

Kraepelin, E. (1893). *Psychiatrie: Ein kurzes lehrbuch fur studirende und aerzte.* Leipzig: Verlag von Ambr. Abel (Arthur Meiner).

Murphy, K. R., & Davidshofer, C. O. (1994). *Psychological testing: Principles and applications.* Englewood Cliffs, NJ: Prentice-Hall.

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

Robins, E., & Guze, S. B. (1970). Establishment of diagnostic validity in psychiatric illness: Its application to schizophrenia. *American Journal of Psychiatry, 126,* 983-987.

Rubinow, D. R., & Roy-Byrne, P. (1984). Premenstrual syndromes: Overview from a methodological perspective. *American Journal of Psychiatry, 141,* 163-172.

Schnurr, P. P., Hurt, S. W., & Stout, A. L. (1994). Consequences of methodological decisions in the diagnosis of late luteal phase dysphoric disorder. In J. H. Gold & S. K. Severino (Eds.), *Premenstrual dysphorias: Myths and realities* (pp. 19-46). Washington, DC: American Psychiatric Press, Inc.

Simon, H. (1978). *Mind and madness in ancient Greece.* Ithaca, NY: Cornell University Press.

Smith, R. M. (1996). A comparison of the Rasch separate calibration and between fit methods of detecting item bias. *Educational and Psychological Measurement, 56,* 403-418.

Wright, B. D., & Linacre, J. M. (1995). BIGSTEPS: Rasch analysis for all two-facet models. Chicago: MESA Press.

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

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

JOURNAL OF OUTCOME MEASUREMENT, *1*(1), 19-33

Constructing Rater and Task Banks for Performance Assessments

George Engelhard, Jr.
Emory University

The purpose of this paper is to present a set of procedures based on Rasch measurement theory for construction an assessment network. An assessment network is defined as a connected system of rater and task banks. Three general classes of data collection designs are presented that can be used to calibrate an assessment network; these are complete, incomplete, and non-linked assessment networks. Carefully constructed assessment networks based on Rasch measurement theory and sound data collection designs provide the opportunity to achieve objective and fair measurements.

Requests for reprints should be sent to Professor George Engelhard, Jr., Division of Educational Studies, North Decatur Building, Emory University, Atlanta, GA 30322

This paper describes a set of procedures for constructing an assessment network composed of a connected system of rater and task banks for large-scale performance assessments. These ideas grew out of work on the development of a large-scale assessment program for measuring writing competence on a high school graduation test. A major goal of this work has been to develop a calibrated set of raters and writing tasks that can be used for the objective measurement of writing competence. In order to accomplish this goal, the focus was on meeting the requirements of objective measurement within the framework of the Rasch model. It is useful to view the calibration of the assessment tasks and the measurement of individuals as separate, although complementary, activities. This approach is congruent with accepted measurement practices; typically, measurement practitioners first calibrate their instruments, and then administer these instruments along with appropriate checks on whether or not each examinee is being assessed objectively and fairly. An assessment network depends on the measurement model selected, as well as the data collection design used to calibrate the facets of the assessment network.

Choppin (1968,1978, 1982) described how item banks can be used to contribute to the improvement of measurement. He defines an item bank as:

> The term 'item bank' should be understood to mean a collection of test items organised and catalogued in a similar way to books in a library. This organising and cataloguing takes account of the content of the test item and also its measurement characteristics (such as difficulty, reliability, validity, etc.). Such items can be readily grouped into tests which will then be properly defined and calibrated measuring instruments" (Choppin, 1978, p. 1).

Based on this definition of an item bank, a task bank can be defined as a calibrated set of prompts whose content and measurement characteristics have been systematically examined and cataloged. In a similar fashion, a rater bank can be defined as a calibrated set of judges whose measurement characteristics have been systematically examined and cataloged. In large-scale performance assessments, it is useful to extend this idea to include networks (Engelhard & Osberg, 1983) with an assessment network defined as a calibrated measurement system composed of rater and task banks. In the language of ANOVA, the crossing of the rater and task banks yields an assessment network that is composed of a variety of assessment components; each assessment component yields an assessment opportunity for an examinee to obtain an observed rating or score. This paper extends the idea of item banks

to include both task and rater banks, as well as the construction of an assessment network composed of a coherent set of banks. In terms of the classification system for linking procedures proposed by Mislevy (1992), the procedures described in this paper reflect calibration more closely than equating.

In the first section of this paper, an extended version of the Rasch model is described that can be used to construct a consistent and coherent performance assessment network. In the next section, illustrative data collection designs that may be used to calibrate an assessment network are described.

A FACETS MODEL FOR WRITING ASSESSMENT

The general model for the assessment of written composition that guides this paper is presented in Figure 1. Ideally, writing competence should be the major variable affecting the observed rating. In practice, when the measurement of writing competence is based directly on student compositions, there are a variety of factors, such as rater and writing task characteristics, that may be viewed as intervening variables. The assessment process should minimize, as much as possible, the effects of these intervening variables on the estimates of writing competence. The situation becomes even more complex when different students are rated by different raters who may vary in severity, and also when different students respond to different writing tasks that may vary in difficulty. The development of rater and writing task banks provides the opportunity to statistically adjust for these differences that may appear when students are not rated by all of the raters on all of the writing tasks, and to obtain fairer and more objective estimates of student competence in writing.

The procedures described here for constructing an assessment network composed of rater and writing task banks are based on a multifaceted version of the Rasch measurement (FACETS) model for ordered response categories developed by Linacre (1989). The FACETS model is an extended version of the Rasch measurement model (Andrich, 1988; Rasch, 1980; Wright & Masters, 1982). The FACETS model is an additive linear model based on a logistic transformation of the observed ratings to a logit scale. Using the terminology of regression analysis, the dependent variable is the logistic transformation of ratios of successive category probabilities (log odds), and the independent variables are the facets. For example, if writing competence was measured with several writing tasks with the compositions rated as pass or fail, then an appropriate Rasch model for this dichotomous data can be written as:

$$\ln [P_{ni1}/P_{ni0}] = \beta_n - \delta_i$$

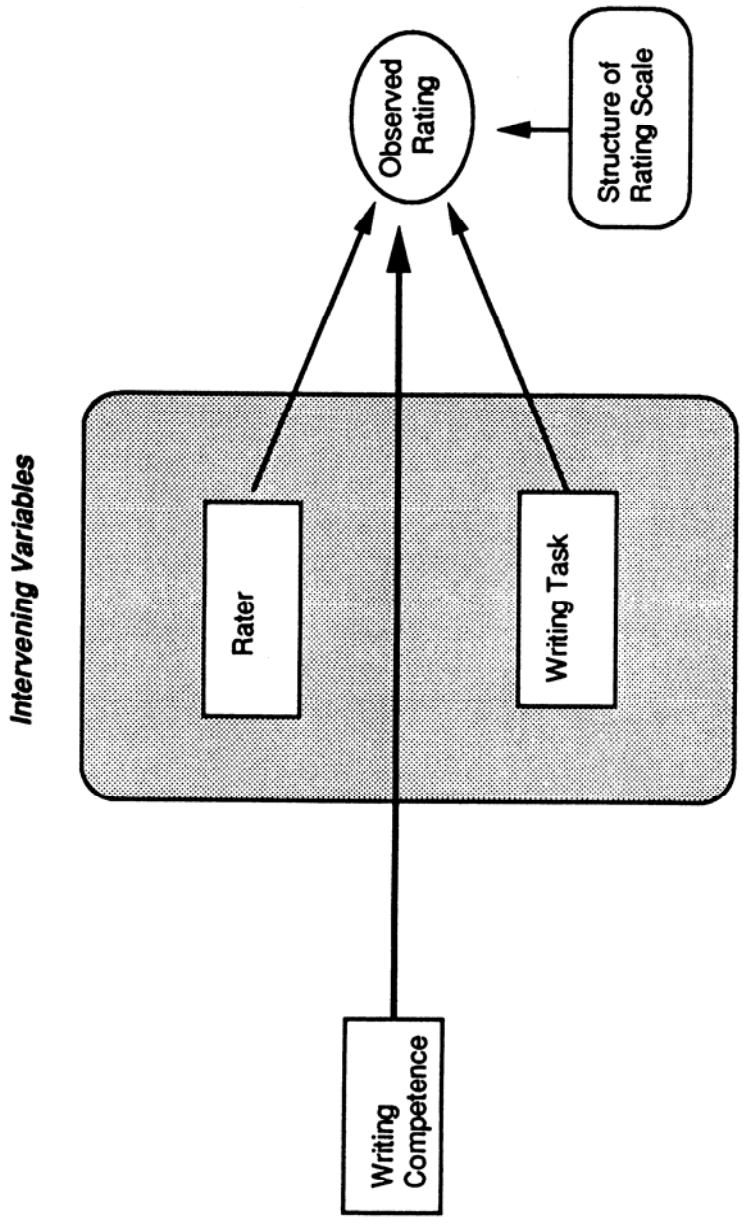


FIGURE 1 Measurement for the assessment of writing competence.

where

P_{ni1} = probability of student n passing (x=1) on writing task i
P_{ni0} = probability of student n failing (x=0) on writing task i
β_n = Writing competence of student n
δ_i = Difficulty of writing task i.

This model has two facets -- student competence and writing task difficulty. This form of the model can be easily extended to deal with rating scale data and multiple facets. The three-facet model (student competence, writing task difficulty, and judge severity) with m +1 rating categories (0, . . ., m) can be written as :

$$\ln[P_{nijk}/P_{nijk-1}] = \beta_n - \delta_i - \lambda_j - \tau_k$$

where

P_{nijk} = probability of student n being rated k on writing task i by rater j
P_{nijk-1} = probability of student n being rated k-1 on writing task i by rater j
β_n = Writing competence of student n
δ_i = Difficulty of writing task i
λ_j = Severity of rater j
τ_k = Difficulty of category k relative to category k-1; i.e., step k

The rating scale parameter, τ_k, which reflects the structure of the four-category rating scale is not labelled as a facet in the model.

The FACETS model is a unidimensional model with a single student competence facet, and a collection of other assessment facets, such as writing tasks and raters. The crossing of these assessment facets defines a set of assessment components that yield multiple ratings for each student. For example, if students responded to two writing tasks and the compositions were rated by three raters, then the assessment network would consist of six assessment components with six observed ratings for each student. The FACETS model is appropriate when the intent of the assessment developers is to sum the ratings from the assessment components in order to produce a total score. As with other Rasch measurement models, the basic assumption of the FACETS model is "that the set of people to be measured, and the set of tasks (items) used to measure them, can each be uniquely ordered in terms

respectively of their competence and difficulty" (Choppin, 1987, p. 111). If the data fit the model and this unique ordering is realized, then a variety of desirable measurement characteristics can be attained. Some of these measurement characteristics are (1) separability of parameters with sufficient statistics for estimating these parameters, (2) invariant estimates of student competence, rater severity and writing task difficulty (this reflects the property of "specific objectivity in Rasch's terminology), and (3) equal-interval scales for the measures. Another way to think about the construction of an assessment network with the FACETS model is to view it as an "equating model" with the raters and writing tasks viewed as analogous to test forms that may vary in difficulty; when different students are rated by different raters on different writing tasks, then it will be necessary to "equate" or statistically adjust for differences in rater severity and writing task difficulty.

Based on the FACETS model presented in Figure 1 and Equation 1, the probability of student n with competence β_n obtaining a rating of x (x = 0, 1, . . . , m) on writing task δ_i from rater λ_j with category step difficulty τ_k is:

$$\pi_{nijx} = \frac{\exp\left[x (\beta_n - \delta_i - \lambda_j) - \sum_{k=0}^{x} \tau_k \right]}{\sum_{s=0}^{m} \exp\left[s (\beta_n - \delta_i - \lambda_j) - \sum_{k=0}^{s} \tau_k \right]}$$

where x=0, ..., m and $\tau_0 \equiv 0$.

Linacre (1989) provides a detailed description of the FACETS model, as well as procedures for estimating the parameters of the model. The fit of rating scale data to the FACETS model can be examined in various ways; Wright and Masters (1982) and Wright and Stone (1979) should be consulted for detailed descriptions of the standardized residuals, the INFIT and OUTFIT statistics, and the reliability of separation index.

DESCRIPTION OF THE DESIGNS

There are a variety of data collection designs that can be used to calibrate raters and writing tasks. In this section, a set of designs are described that

raters and writing tasks. In this section, a set of designs are described that illustrate many of the data collection issues that need to be considered in the construction of rater and writing task banks. An attempt has been made to construct a bridge between the widely accepted language used with equating traditional multiple-choice tests with several forms (Andrich, 1988; Petersen, Kolen, & Hoover, 1989), and the language used with calibrating IRT models (Hambleton, Swaminathan, & Rogers, 1991; Linacre, 1989; Wright & Stone, 1979). The measurement situation used to illustrate the designs is based on two writing tasks, three raters, and ten examinees; extensions of these designs and basic principles to assessment networks with more than three facets are straightforward. Operational designs for calibrating writing tasks and raters would be based on many more examinees and usually more raters. Examinees can be viewed as replications within each cell of the design. Increasing the number of examinees within a cell would result in a concomitant decrease in the standard error for any estimates that included that cell. There are three general categories of designs that can be used for linking assessment components into a consistent and coherent network. These categories are complete, incomplete, and non-linked assessment networks.

Before describing the designs, it is useful to define a few terms. *Facets* are the separate dimensions used in the assessment network. In the language of analysis of variance, facets are factors. Facets are composed of individual *elements* that vary in difficulty. The difficulty of an element defines its location on the latent variable or construct that the assessment network is designed to measure. For example, each writing task is an element within the writing-task facet, and each rater is an element within the rater facet. It should be noted that the examinee is a facet in this model, while in Generalizability Theory examinees are not considered a "facet" (Shavelson and Webb, 1991). When rater and writing task facets are crossed, the cells within the design are called *assessment components;* each assessment component yields an assessment opportunity for the examinee to obtain an observed rating that depends on the difficulty of the elements from each facet that combine to define that cell. The assessment components obtained from crossing several facets combine to define an overall *assessment network.*

Complete assessment networks consist of completely crossed designs with examinees receiving observed scores on all of the assessment components. Examples of these designs are shown in Table 1. These completely crossed designs are the simplest, but also the most expensive, data collection designs.

TABLE 1
Data Collection Designs for Complete Assessment Networks

			Examinee									
Assessment Component	Rater	Task	1	2	3	4	5	6	7	8	9	10
1. Two-Facet Design (task x examinee)												
1		1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2		2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2. Two-Facet Design (rater x examinee)												
1	1		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	2		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	3		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3. Three-Facet Design (rater x task x examinee)												
1	1	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	2	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	3	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	1	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
5	2	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6	3	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note. These designs are generalizations of Single-Group Designs (Petersen, Kolen, & Hoover, 1989). The designs are represented here with 10 examinees. Operational designs would require more examinees. A ✓ indicates that a rating is obtained for the examinee on this assessment component, otherwise a rating is not obtained.

The connectedness of complete assessment networks is presented graphically in the first column of Figure 2. The circles represent the assessment components, and the lines indicate that examinee data is available that provides for the direct estimation of a link between all of the assessment components included in the overall assessment network. In the cases where there are two numbers within a circle, the first number indexes the tasks, and the second number indexes the raters. In practice, it would be desirable to randomize the order of presentation of the writing tasks in order to minimize the effects of extraneous factors, such as learning, fatigue and practice. Context effects may also influence the rating behavior of the raters, and the order of the presentation of the compositions to the raters should also be randomized. The number of assessment components for the Two-Facet Designs (task x examinee and rater x examinee) match the number of elements (tasks or raters) in the design. For the Three-Facet Design, the number of assessment components reflects the product of the number of raters times the number of writing tasks ($3 \times 2 = 6$). These designs for constructing complete assessment networks are generalizations of the Single-Group and Counterbalanced Random Groups Designs described by Petersen, Kolen, and Hoover (1989).

Incomplete assessment networks consist of designs in which examinees do not have scores on all of the assessment components, and systematic links have to be created in order to yield a connected network of assessment components. When developing a calibrated assessment network, there are a variety of practical considerations that rule out the construction of complete assessment networks. Carefully designed incomplete assessment networks can be used to obtain reliable and valid links both within and between facets that are less costly in terms of examinee time and rater salaries. Examples of these types of designs are shown in Table 2.

For two-facet designs (task x examinee or rater x examinee), it is possible to calibrate each facet through common examinees or through an anchor facet (anchor tasks or anchor raters); it is also possible to anchor rating steps. The term "anchor" simply refers to the practice of fixing or presetting the calibrations (scale values) of some or all of the elements within a facet based on prior information. The number of assessment components and the number of observed ratings obtained for each examinee are not the same in an incomplete assessment network. The connectedness of incomplete assessment networks is presented graphically in the second column of Figure 2. For incomplete assessment networks, all of the

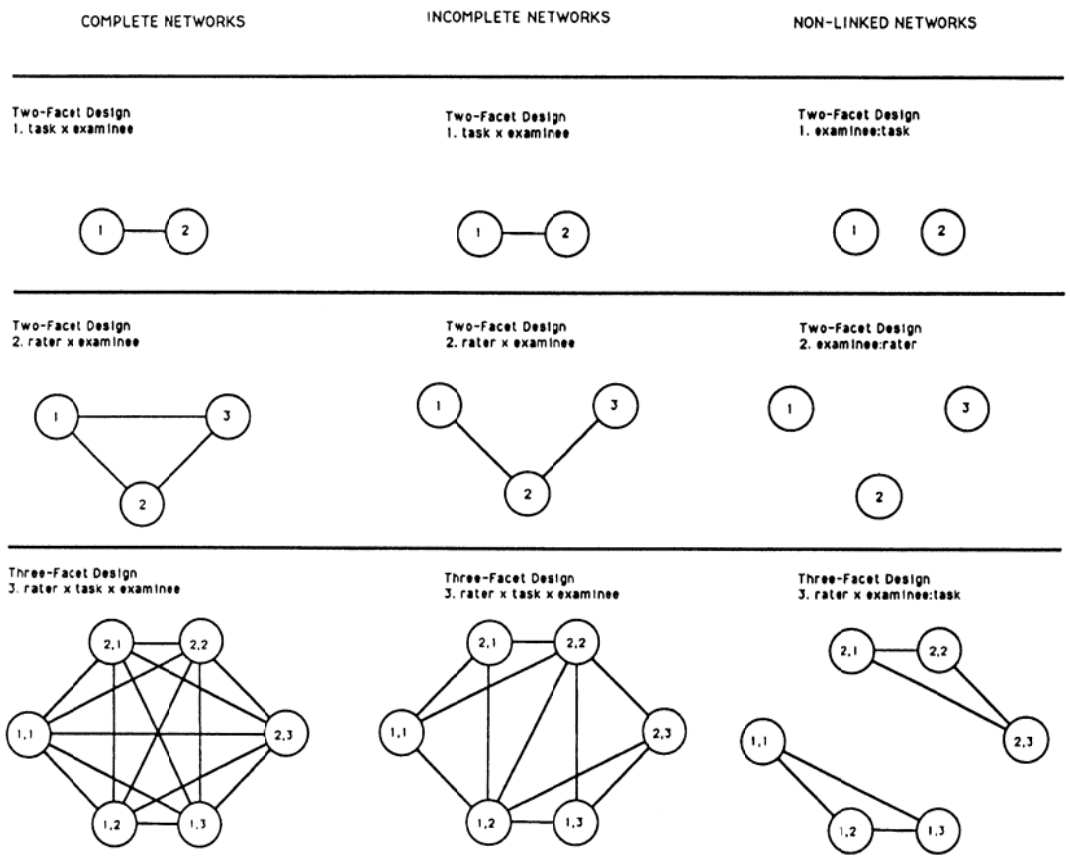


FIGURE 2 Diagrams of Data Collection Designs.

TABLE 2
Data Collection Designs for Incomplete Assessment Networks

Assessment Component	Rater	Task	Examinee 1	2	3	4	5	6	7	8	9	10
1. Two-Facet Design with Common - Examinee Design (task x examinee)												
1		1	✓	✓	✓	✓	✓	✓	✓			
2		2				✓	✓	✓	✓	✓	✓	✓
2. Two-Facet with Anchor - Rater Design (rater x examinee)												
1	1		✓	✓	✓	✓	✓					
2	2		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	3							✓	✓	✓	✓	✓
3. Three-Facet with Anchor - Rater Design (rater x task x examinee)												
1	1	1	✓	✓	✓	✓	✓					
2	1	2	✓	✓	✓	✓	✓					
3	2	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	2	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
5	3	1						✓	✓	✓	✓	✓
6	3	2						✓	✓	✓	✓	✓

Note. These designs are generalizations of Anchor - Test Designs (Petersen, Kolen, & Hoover, 1989). The designs are represented here with 10 examinees. Operational designs would require more examinees. A ✓ indicates that a rating is obtained for the examinee on this assessment component, otherwise a rating is not obtained.

assessment components are linked together, although there are fewer links. The construction of connected incomplete assessment networks is complex, and there are many choices for acceptable designs. The data collection designs used to construct incomplete assessment networks are examples of Balanced Incomplete Block (BIB) and Partially Balanced Incomplete Block (PBIB) designs with block sizes of at least two. Thus, there are a plethora of designs that can be considered (John,1980; Kirk,1968). BIB and PBIB designs make it possible to estimate "main effects," but the situation becomes more complicated when bias analyses and differential facet functioning based on interactions among the facets need to be explored. If systematic links are not built into the data collection design, then non-linked assessment networks may result; Weeks and Williams (1964) have described a straightforward procedure for identifying linked assessment networks, and this procedure is used in the FACETS computer program to check for connectedness (Linacre & Wright, 1992). Many of these issues also appear in the literature on paired comparisons (David, 1988). These designs are generalizations of the Anchor-Test Designs described in Petersen, Kolen, and Hoover (1989).

Non-linked assessment networks are designs in which examinees do not have scores on all of the assessment components, and some systematic links among the assessment components are missing. Examples of these types of designs are shown in Table 3. The lack of connectedness in non-linked assessment networks is presented graphically in the third column of Figure 2. These designs lead to assessment networks that break into two or more disconnected networks of assessment components depending on the nesting structure of the data collection design. In the language of analysis of variance, nested facets are nested in a second facet if each element of the first facet (nested facet) appears in only one element of the second facet. For example, if 5 examinees respond to Task 1 and 5 different examinees respond to Task 2, then examinees are nested within task because each examinee only appears in one element of the second facet (either Task 1 or Task 2); no examinee responds to both tasks. These designs have many weaknesses, and some measurement professionals might even question including these designs or even calling them "networks." The quality of the network depends on how well the "equivalent" groups are defined. The nesting structure makes it impossible to *directly* calibrate all of the assessment components, and additional assumptions are required to *indirectly* connect the disconnected assessment components. For example, if the writing tasks are not directly linked, then it is not possible to eliminate the potential influences of the particular examinees used to calibrate the

TABLE 3
Data Collection Designs for Non-linked Assessment Networks

			Examinee									
Assessment Component	Rater	Task	1	2	3	4	5	6	7	8	9	10
1. Two-Facet Design (examinee: task)												
1		1	✓	✓	✓	✓	✓					
2		2						✓	✓	✓	✓	✓
2. Two-Facet Design (examinee: rater)												
1	1		✓	✓	✓							
2	2					✓	✓	✓	✓			
3	3									✓	✓	✓
3. Three-Facet Design (rater x examinee: task)												
1	1	1	✓	✓	✓	✓	✓					
2	2	1	✓	✓	✓	✓	✓					
3	3	1	✓	✓	✓	✓	✓					
4	1	2						✓	✓	✓	✓	✓
5	2	2						✓	✓	✓	✓	✓
6	3	2						✓	✓	✓	✓	✓

Note. These designs are generalizations of Equivalent-Groups Designs (Petersen, Kolen, & Hoover, 1989) the designs are represented here with 10 examinees. Operational designs would require more examinees. A ✓ indicates that a rating is obtained for the examinee on this assessment component, otherwise a rating is not obtained.

assessment network without additional assumptions. These designs for constructing non-linked assessment networks are generalizations of the Equivalent-Groups Designs described by Petersen, Kolen, and Hoover (1989).

DISCUSSION

This paper focused on the description of procedures for constructing an assessment network composed of rater and task banks. Item banks have provided a useful framework for solving a variety of measurement problems encountered with selected-response items (Wright & Bell, 1984). It is expected that assessment networks composed of rater and task banks will provide a similar framework for improving measurement practices with constructed-response items and other types of performance assessments. The work presented in this paper is guided by the requirements of objective Rasch measurement. It is also guided by the view that a systematic set of procedures and data collection designs should be used to provide as much control as possible over the quality of the data collected.

Three general categories of designs for linking assessment components into a consistent and coherent network were described. These are complete, incomplete, and non-linked assessment networks. Data collected based on these designs can be analyzed with the FACETS model described in this paper using the FACETS computer program (Linacre & Wright, 1994). Carefully constructed assessment networks based on sound data collection designs provide the opportunity to achieve objective and fair measurements within complex assessment systems with multiple facets.

ACKNOWLEDGEMENTS

I would like to thank Belita Gordon, Steve Gabrielson, and David Curtin for their contributions to the ideas presented in this paper. In addition, I would like to thank two anonymous reviewers for their helpful comments on an earlier version of the manuscript.

REFERENCES

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

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

Choppin, B. (1978). *Item banking and the monitoring of achievement.* (Research

in Progress Series No. 1). Slough: National Foundation for Educational Research.

Choppin, B. (1982). The use of latent trait models in the measurement of cognitive abilities and skills. In D. Spearritt (Ed.), *The improvement of measurement in education and psychology* (pp. 41-63). Melbourne: Australian Council for Educational Research.

Choppin, B. (1987). The Rasch model for item analysis. In D.L. McArthur (Ed.), *Alternative approaches to the assessment of achievement* (pp. 99-127). Norwell, MA: Kluwer Academic Publishers.

David, H.A. (1988). *The method of paired comparisons.* London: Charles Griffin & Company Limited.

Engelhard, G, & Osberg, D.W. (1983). Constructing a test network with a Rasch measurement model. *Applied Psychological Measurement, 7*, 283-294.

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

John, P.W.M. (1980). *Incomplete block designs.* New York: Marcel Dekker, Inc.

Kirk, R.E. (1968). *Experimental design: Procedures for the behavioral sciences.* Belmont, CA: Brooks/Cole Publishing Company.

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

Linacre, J. M., & Wright, B.D. (1994). *A user's guide to FACETS: Rasch measurement computer program.* Chicago: MESA Press.

Mislevy, R.J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: ETS, Policy Information Center.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, Norming, and equating. In R.L. Linn (Ed.), *Educational Measurement*, Third Edition (pp. 221-262). New York: American Council on Education and Macmillan Publishing Company.

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

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications, Inc.

Weeks, D.L., & Williams, D.R. (1964). A note on the determination of connectedness in an n-way cross classification. *Technometrics,* (3), 319-324.

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

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

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

JOURNAL OF OUTCOME MEASUREMENT, 1(1), 34-55

Development of a Scale to Assess Concern about Falling and Applications to Treatment Programs

Michelle M. Lusardi
University of Connecticut

Everett V. Smith Jr.
University of Oklahoma

This study used Rasch methodology to pursue three goals. First, we sought to demonstrate the psychometric limitations of the Falls Efficacy Scale (Tinetti, Richman, & Powell, 1990). Second, we addressed these limitations using a simultaneous calibration of the Falls Efficacy Scale and Mobility Efficacy Scale items. Third, we review previous explorations of the self-efficacy construct in relationship to health behaviors and discuss a possible treatment program based on the simultaneous calibrated items and Social Cognitive Theory. Results indicate that responses from the Falls Efficacy Scale fail to assess the higher ends of the self-efficacy continuum. Simultaneous calibration of items improved this lack of scale definition. This initial work in assessing self-efficacy perceptions provides a theoretical framework for planning treatment programs that may be more cost effective than collecting performance measures.

Authors are listed alphabetically as each contributed equally to this project.
Requests for reprints should be sent to Everett Smith at The University of Oklahoma, Department of Educational Psychology, Collings Hall, Norman, Oklahoma, 73019

Concern about functional mobility and safety has prompted intense research on physical infirmity and risk of falling among frail elderly persons (Campbell, Reinken, Allan, & Martinez, 1981; Magaziner, Cadigan, Hebel, & Parry, 1988; Manton, 1988; Prudham & Evans, 1981; Tinetti, Speechley, & Ginter, 1988). The sequelae of accidents, primarily falls, ranks in the top ten causes of death in the elderly (Baker & Harvey, 1985; Tinetti, 1992; Wild, Nayak, & Isaacs, 1981). Although the majority of falls do not lead to death, fall-related morbidity is significant (Evans, 1988; Melton & Riggs, 1985). Prospective studies of community living elderly indicate that injuries (fracture of hip, wrist, rib, or vertebras, head injury, hemarthrosis, laceration, abrasion, and ecchymosis) occur in 25% of falls, limiting functional mobility and compromising performance of activities of daily living (DeVito et al., 1988; Jette, Branch, & Berlin, 1990). Speechley and Tinetti (1991) report that as many as 40% of elderly who fall, whether injured or uninjured, are unable to rise from the floor without assistance. Falling and fall related injury are serious health problems for older persons and their care-givers.

Epidemiological studies have consistently found an incident of falls of 30 to 50% in community living elderly persons (Campbell, Borrie, & Spears, 1989; Nevitt, Cummings, Kidd, & Black, 1989; Perry, 1982; Reinsch, MacRae, Lachenbruch, & Tobis, 1992; Tinetti et al., 1988). Among institutionalized older persons, the incidence ranges from 50 to 75% (Tinetti, Williams, & Mayewski, 1986). This incidence, combined with the risk of resultant injury and impairment, has been a major stimulus for investigation of biomedical, environmental, and age-related physiological factors which contribute to frailty, risk of falling, and fall-related injury.

We now understand falling to be a consequence of the interaction of cumulative age-related physiological changes and pathological impairments which compromise dynamic postural responses (Chandler & Duncan, 1993; Hindmarsh & Estes, 1989). Currently, there is much emphasis on interventions (e.g., strength training, environmental modifications) designed to improve postural response and reduce risk of falling (Wolf, Kutmer, Green, & McNeely, 1993; Wolfson et al., 1993). Although early results suggest success in reducing fall episodes, an increasing number of investigators are identifying a problematic "post-falls" fear syndrome of self-imposed activity restriction, often drastically out of proportion to the consequences of the fall itself (Bhala, O'Donnell, & Thoppil, 1982; Downton & Andrews, 1990; Holiday, 1992; Maki, Holiday, & Topper, 1991; Murphy & Isaacs, 1982; Tideiksaar & Silverton, 1989; Walker & Howland, 1991). Studies of community living elderly have identified "fear of falling" and restriction of ac-

tivity even among older persons who have not fallen (Speechley & Tinetti, 1991; Vellas, Cayla, Bocquet, dePemille, & Albarede, 1987). The consequences of reduced activity are significant: loss of strength, stamina, and flexibility quickly lead to functional impairment and decline, further compromising postural responses (Fiatarone et al., 1990), thus increasing risk of falling.

To date, the factors which contribute to "fear of falling" have only been loosely described, without adequate definition or systematic investigation. Although clinical observations of health professionals support the relationship between fear of falling and inactivity, we do not understand the circumstances, concerns, or beliefs about aging which contribute, appropriately or inappropriately, to fear of falling and resulting restrictions of activity. The psychological characteristics of older persons who are fearful of falling have not yet been investigated, although trait anxiety and depression have been implicated (Holiday, 1992; Maki et al., 1991; Tinetti, Richman, & Powell, 1990). Interventions aimed at minimizing risk of falls and reducing fall occurrence will be most efficacious if both physical and psychological risk factors are addressed. It is important to explore the physical and psychological characteristics of older persons who are afraid of falling.

The Falls Efficacy Scale (FES; Tinetti et al., 1990) was the first instrument developed to measure fear of falling in older persons using a social cognitive model of self-efficacy. Concerns about the content validity, and a ceiling effect in scoring prompted the authors to develop additional items, the Mobility Efficacy Scale (MES) modeled on the FES. This paper will describe the social cognitive model of self-efficacy as it relates to rehabilitation, describe the instrument development for the MES, and evaluate the construct validity of the FES and combined FES/MES responses using Rasch modeling. The paper will conclude with a discussion of the potential contributions of the assessment of self-efficacy based on Rasch modeling to intervention planning and outcomes evaluation in rehabilitation of older adults.

MODELS

Social Cognitive Theory

Social Cognitive Theory suggests that people learn through direct and indirect observation and vicarious reinforcement. This allows people to exercise control over their thoughts, feelings, and action. A major component of this theory is a construct referred to as self-efficacy: an individual's judg-

ment about being able to perform a specific behavior. Self-efficacy is thought to mediate between knowledge and behavior. It is a persons "I can" or "I cannot do" belief and is not concerned with the skills one has, but with the judgments of what one can do with the skills one possesses. That is, knowledge itself is not enough to motivate behavior, moderate to high efficacy expectations (or extremely good incentives) need to be present for an individual to engage in a particular behavior. Self-efficacy helps explain why people's behavior may differ dramatically even when they share similar knowledge and skill levels. These characteristics of the self-efficacy construct are especially relevant for the elderly, all of whom have similar knowledge via personal experiences to perform the various behaviors found in the FES and MES, but whose differing performance may be due to differences in their sense of self-efficacy to engage in the activities. Compared to those with low self-efficacy, people highly self-efficacious have more perseverance, set more challenging goals, continue in the face of difficult barriers and occasional failures, and will attribute success to ability and effort and failure to a lack of effort (Bandura, 1986). Their high self-efficacy motivates behavior that produces accomplishments. In contrast, those with low self-efficacy shy away from difficult tasks, lack effort, give up easily when faced with difficult tasks, are easily distracted by thoughts of personal deficiencies, and attribute success to luck or ease of task and failure to lack of ability. The way one perceives oneself will affect a willingness to approach a task and put forth maximum effort. To change an individual's non-productive behavior, it is necessary to change or raise their self-efficacy.

Most evaluations of medical or educational programs rely on observable behavior while overlooking other important constructs that have been demonstrated to be predictive of current and future performance, namely self-efficacy. Self-efficacy influences persistence and motivation, both of which are important outcomes to any intervention or treatment program (physical or educational). In conjunction with performance data, self-efficacy measures can serve as an important part of program planning and evaluation, indicating behaviors individuals do not possess sufficient confidence in their ability to perform, either prior to, during, or after a treatment program. These self-assessed weaknesses can suggest a more efficient course of treatment for a present or future program. The limitation of using only performance measures is that even though it may appear individuals have mastered a given behavior, performance measures alone give no indi-

cation as to whether these behaviors will be successfully attempted in the future. Treatment programs are limited if alteration of skills is achieved but they do not endow the patient with confidence and motivation to engage in the behavior in the future.

The influence of self-efficacy beliefs on performance of activities of daily living and functional impairment in older adults is receiving increased attention by clinical researchers. There is a strong relationship between self-efficacy beliefs and health-promoting behaviors intended to prevent functional decline, such as initiating participation in exercise program or continuing to exercise in the face of stressors (McAuley, Lox, & Duncan, 1993). In a recent study of older adults with osteoarthritis, self-efficacy beliefs were found to be as powerful a predictor of speed of performance during stair climbing activities as the presence of pain (Rejeski, Craven, Wittinger, McFarlance, & Shumaker, 1996). Self-efficacy beliefs for walking have been found to be one of the important predictors of survival rates in patients with chronic obstructive lung disease (Kaplan, Ries, Prewitt, & Eakin, 1994). Mendes de Leon and colleagues (1996) have found low self-efficacy to be predictive of functional decline over 18 months in a large sample of community living older adults, while high self-efficacy appeared to have a protective effect between physical capacity and functional decline. In addition, evidence is accumulating that interventions targeted at enhancing self-efficacy beliefs impact on behaviors and functional performance. Studies of patients with rhuematoid arthritis (Buescher et al., 1991; O'Leary, Shoor, Lorig, & Holman, 1988) and with osteoarthritis (Lorig & Holman, 1993) have demonstrated that "efficacy" interventions modified perception of pain, increased willingness to participate in activities previously associated with pain, and increased overall levels of physical activity.

Tinetti, Mendes de Leon, Doucette, & Baker (1994) investigated the relationship between fear of falling, falls efficacy, and functional status in older adults. In both fallers and non-fallers, low self-efficacy beliefs measured on the FES were associated with poor performance on a variety of physical performance measures. Myers et al. (1996) explored the relationship between functional status and balance self-efficacy using comparison groups of elders with "high" and "low" mobility. Functional status was measured by dynamic posturography and gait analysis. Balance self-efficacy was measured using the FES and the Activities Specific Balance Confidence Scale (ABC; Powell & Myers, 1995). Subjects with high mobility and with high levels of balance self-efficacy were more stable on posturography. Subjects with low FES and ABC scores were more likely

to avoid activities to reduce risk of falling. Both Tinetti and Myers suggest that prevention and rehabilitation interventions targeting both physical performance and confidence levels are important when working with older adults at risk of functional decline and falling.

Self-efficacy measures are easy to construct and generally show strong estimates of internal consistency and factorial validity (Froman & Owen, 1991). These self-report measures are rapid to administer and non-threatening. This is in contrast to performance data, which may take considerable time to construct, administer, and score. The specificity of efficacy beliefs, however, limits the use of self-efficacy scales across contexts or situations. This implies that self-efficacy scale items must be specifically targeted at the behavioral goals of the planned interventions.

Item Response Theory

Program planning and evaluation typically use aggregate data. As a result, individual variability is lost. Additionally, the same total score may be reached through numerous combinations of responses at the item level. There is a need to be able to locate individuals who are different from the group both prior to treatment and upon completion of the program. Aggregate scores based on responses to a Likert format are ordinal. This makes valid comparisons among or between individuals or items prior to or following treatment difficult as equal score differences between different pairs of points do not imply equal amounts of the construct under investigation. Another limitation with the comparison of raw scores is that these comparisons will always depend on which items are administered and if norms are used, which sample of subjects provided the norms.

Rasch measurement models overcome these limitations. Rasch models are mathematical models that specify unidimensionality and additivity. Unidimensionality means that all items measure a single construct. Additivity refers to the properties of the measurement units, which are the same size (i.e. interval) over the entire continuum. These units are called logits (logarithm of odds) and are a linear function of the probability of responding to a given category on a Likert scale for a person of a given ability.

Rasch models also estimate item calibrations independently of the sample employed and person measures independently of the items used. The degree to which these properties hold depends on how closely the data fit the model. Once the parameters of a Rasch model are estimated, they are used to compute expected (predicted) response patterns on each item.

Fit statistics are then derived from a comparison of the expected patterns and the observed patterns. These fit statistics are used as a measure of the validity of the model-data fit and as a diagnosis of individual idiosyncrasy.

Item fit statistics are used to verify the internal validly of the items in contributing to a unitary scale. The model requires that an item have a greater probability of yielding a higher rating for persons with higher ability than for persons with lower ability. Those items identified as not fitting the Rasch model need to be examined and either revised or eliminated. Such an item may not be related to the rest of the scale (e.g., assessing a concept other than that shared by the remaining items).

Person fit statistics measure the extent to which a person's pattern of responses to the items correspond to that predicted by model. A valid response, as specified by the model, dictates that a person of a given ability have a greater probability of providing a higher rating on easier items than on more difficult items. Persons identified as misfitting may not be from the targeted sample or the content of the assessment may not be appropriate for the given person.

BIGSTEPS (Linacre & Wright, 1995) provides two types of fit statistics for persons and items: Infit, which is sensitive to unexpected responses to items near a person's ability level, and Outfit, which is sensitive to aberrant behavior on items far from a person's ability level. When reported as standardized values, these fit statistics have an expected value of zero and a standard deviation of one. Values less than zero suggest a lack of variability in the data. Values greater than zero are indicative of excessive variability. A reasonable range for both types of fit statistics is -2 to 2. Items or persons with fit statistics outside this range need to be evaluated in order to determine the possible cause of the misfit. Standardized residuals, with an expected value of zero and standard deviation of one, may be helpful for determining why particular items or individuals fail to follow the Rasch model. Negative residuals indicate unexpected low responses while positive residuals indicate unexpected high responses. Since one of the purposes of this investigation is scale development, the primary focus will be given to item fit statistics.

While fit statistics address validity in the context of Rasch modeling, standard errors associated with each item calibration and person ability estimate provide evidence for reliability. These errors can be used to describe the range (i.e. confidence intervals) within which each item's "true" difficulty or person's "true" ability falls. These errors can also be used to determine strata: regions of the scale whose centers are separated by logit distances greater than can

be accounted for by measurement error. Mathematically, strata are the quotient of four times the separation index plus one (4G +1) divided by three. It has been suggested that a scale must reach out to at least two item difficulty strata to be useful for scale definition (Kilgore, Fisher, Silverstein, Harley, & Harvey, 1993).

METHOD

Sample

The population targeted for this study was community living older adults. Three criteria were used to determine eligibility:

1. Age of at least 65 years. Exceptions were made for four sub jects who had functional mobility impairment or a history of repeated falls.
2. Ability to ambulate, with or without an assistive device. This was assessed by self-report and direct observation.
3. Cognitive function adequate to complete a pencil and paper questionnaire. Adequacy was defined as four or fewer errors on the Short Portable Mental Status Questionnaire (Pfeiffer, 1975).

Subjects were recruited from senior housing complexes, senior centers, and area churches in five Connecticut communities. A short recruiting presentation describing the purpose of the study, eligibility criteria, time requirements, and benefits of participation was made at each site. Initially 131 women and 12 men indicated willingness to participate. On follow-up, 38 women and four men changed their minds about participating, citing chronic or acute illness, transportation problems, or schedule conflicts. Two subjects who failed to respond to the pencil and paper tasks were removed from the sample. This left 92 women and 8 men who completed all phases of the study. The very small number of men in the sample precluded statistical examination of gender differences, and a decision was made to exclude this subsample from further analysis. This decision is also supported by the literature on mobility impairment and falls in later life: women are more likely to be living alone (Magaziner et al, 1988), have greater difficulty with mobility impairment and instrumental activities of daily living (Jette, Branch, & Berlin, 1990), and are more likely to experience recurrent falls and fall related injury (Mendes de Leon et al., 1996). Analysis

was performed on a final sample of 92 older women. Their mean age was 76.12 years (*SD* = 6.75) with a range of 59 to 91. Sixty of these women lived alone. Ninety-one were Caucasian. One was African-American. Thirty-five had a high school education or less, 18 had a college education, and seven had attended graduate school.

Subjects were evaluated in a "balance clinic" which screened for biomedical and physical risk factors for falls and evaluated mobility and dynamic balance. One purpose of the clinic was to provide researchers with a baseline of performance for interpreting subject's efficacy appraisals. Following the balance clinic, appointments were made for in-home follow-up interviews. All interviews were performed by a single trained interviewer, and included the Falls Efficacy (Tinetti et al., 1990) and Mobility Efficacy scales.

Instruments

At the time of the study, the FES was the only instrument in the literature concerned with fear of falling (Tinetti et al., 1990). Fear of falling is defined as significant worry, concern, or anxiety about the potential for falling and negative consequences of a fall. Subjects indicate their efficacy in avoiding falls on a 4-point Likert scale. Tinetti et al. (1990) report evidence of discriminate validity (those avoiding activities because of self-reported fear scored significantly higher on the FES than those reporting no fear) and test-retest reliability of r = .71.

The FES items were developed by a panel of health professionals (MD, RN, PT, OT) involved in caring for frail older persons in hospital, long term care, and home care settings. FES items reflect those behaviors necessary for safety and functional independence in basic activities of daily living. In previous research involving community samples, responses to FES items are skewed toward high self-efficacy. This type of ceiling effect occurs when many items of low difficulty are included in a scale (Borg & Gall, 1989). A review of item stems suggests that these very basic activities of daily living may not accurately reflect the functional tasks which challenge postural stability of older adults, indicating that a limited portion of the self-efficacy continuum was being assessed. A more accurate way of assessing fear of falling may be to ask older persons themselves, as "content experts" (Gable & Wolf, 1993) on mobility issues in later life, about the activities they perceive as having some degree of postural challenge or risk

of falling.

The MES was developed to include a variety of activities more challenging to postural control than the activities of daily living in the FES. Item stems were developed based on focused discussions with three groups of community living older persons (total n=31). Age of group members ranged from 69 to 92 years. Groups included both men and women. There was much diversity in functional ability within the group: ranging from active, healthy individuals, to those with moderate to severe limitations from arthritis, visual impairment, and hearing loss. Almost 1/3 of the group reported a fall in the previous six months, and almost all knew of someone who had sustained an injury by a fall.

The degree to which independent "content expert" groups generate similar item stems increases confidence that a measure will assess what it intends (Gable & Wolf, 1993). A multi-step process was used in the development of the MES item stems. Group 1 (n=3) generated the first five items. Group 2 (n=3) agreed that these five were relevant, and added three additional items. A larger Group 3 (n=25) independently generated a list of activities. They listed 7 of the 8 previously described activities, agreed that the 8th was relevant and added two additional items. The resulting items include motor activities which require more complex postural control than many of the FES items. This increasing level of difficulty would potentially address the "ceiling effect" encountered in FES responses. The contents of the FES and MES are in the Appendix.

A common rating scale was adopted for all 20 items. For each item, participants responded to the question "How concerned (about your ability) are you that you might fall when you are……". Self-reported responses were recorded on a 4-point Likert scale with the points labeled "not at all concerned" (1), "a little concerned" (2), "fairly concerned" (3), and "very concerned" (4). Prior to analyses responses were reversed so that higher scores represented higher self-efficacy for avoiding falls. Total testing time for the FES and MES did not exceed 10 minutes.

Analyses

The Rasch rating scale model (Wright & Masters, 1982) was employed for all analyses. Using this model, a self-efficacy parameter for each person, a set of scoring category threshold parameters common to each item, and item challenge parameters were estimated.

Data from the FES were analyzed first in order to verify the hypothesis that items from this assessment span only a lower portion of the self-efficacy continuum. Data from the FES and MES were then calibrated simultaneously. For each assessment, the respective item calibrations were anchored at the simultaneous calibrated values and then used to generate a measure for each person. If the FES and MES measure the same construct, the measures from the two separate calibrations and the simultaneous calibration should be highly correlated.

RESULTS

Falls Efficacy Scale

The analysis of the FES data confirmed the hypothesis that the FES assesses a lower portion of the self-efficacy continuum. The range of item calibrations was from 1.59 (FES4) to -1.31 (FES2). The item reliability index, an indicator of the spread of item difficulties along the self-efficacy continuum, of .95 translates into 6.43 statistically distinct item challenge strata the persons have distinguished. The person reliability index, an indicator of the spread of person self-efficacy estimates along the self-efficacy continuum, was .67. This reliability index translates into 2.24 statistically distinct self-efficacy strata distinguished by the FES items.

Although several item challenge strata were distinguished and approximately two levels of self-efficacy were noted, the suitability of the FES scale for this sample is questionable. The average person measure of 1.89 logits (*SD* = 1.11) exceeds even the highest item calibration of 1.59 (FES4) indicating that scale is failing to assess the higher ends of the self-efficacy continuum.

Simultaneous Calibration

Given the lack of FES scale definition at the higher ends of the self-efficacy continuum, the ten MES and ten FES items were calibrated simultaneously in order to investigate improvements in scale definition when all items were in the same unit of measurement. The initial calibration identified the FES item "Reaching into cabinets or closets" as problematic, with an estimated measure of -.02 and a standardized INFIT value of 2.3 and a standardized OUTFIT value of 1.4. Examination of standardized residuals revealed several negative values. This implies that there were several highly efficacious people providing unexpected low responses, indicating a spo-

radic lack of confidence in performing this activity. Perhaps "Reaching into cabinets or closets" for this group of individuals measures a construct other than that defined by the remaining items. This item was subsequently eliminated from the item pool as it contributed to discrimination only toward the lower end of the self-efficacy continuum and the elimination of this item did not produce any changes in the person and item reliability indices upon re-calibration.

Re-calibration of the remaining 19 items yielded a person reliability index of .89. This reliability index translates into 4.08 statistically distinct self-efficacy strata distinguished by the items. The item reliability of .98 translates into 9.81 statistically distinct item challenge strata that the persons distinguished. Both these indices demonstrate substantial improvement in scale definition over the results produced by the FES item calibration. As seen in Figure 2, with the inclusion of the MES items the range of the item calibrations (2.93 to -1.87, SD = 1.30) now spans much of the distribution of person self-efficacy estimates (range 4.49 to -1.38, M = 1.35, SD = 1.21) leading to increased measurement precision at the higher ends of the self-efficacy continuum.

Separate FES and MES measures were produced for each person using the simultaneous item calibrations as anchor values. The person measures from the separate calibration of the FES items correlated .89 ($p<.0001$) with the simultaneous calibrated person measures. For the person estimates from the MES this correlation was .97 ($p<.0001$). The separate estimates of person self-efficacy measures correlated .77 ($p<.0001$). These correlations demonstrate the FES and MES person estimates to be moderately correlated with each other and highly correlated with the simultaneously calibrated person estimates, indicating a common underlying construct. Figure 1 shows the scatterplots for the correlations between the FES and simultaneous and the MES and simultaneous person estimates. As expected, the correlation between the FES and simultaneous person estimates is precise only in the lower portion of the self-efficacy continuum, with large fluctuations becoming apparent as one moves to higher levels of self-efficacy. This is in contrast to the correlation between the MES and the simultaneous person estimates in which the measurement precision is consistent throughout the continuum.

The construct validity of responses in Rasch measurement is addressed by examining the sequence and calibrations of items. Items should define a hierarchical scale of activities that represent a unidimensional concept. The variable map in Figure 2 displays the hierarchical nature of this scale. Per-

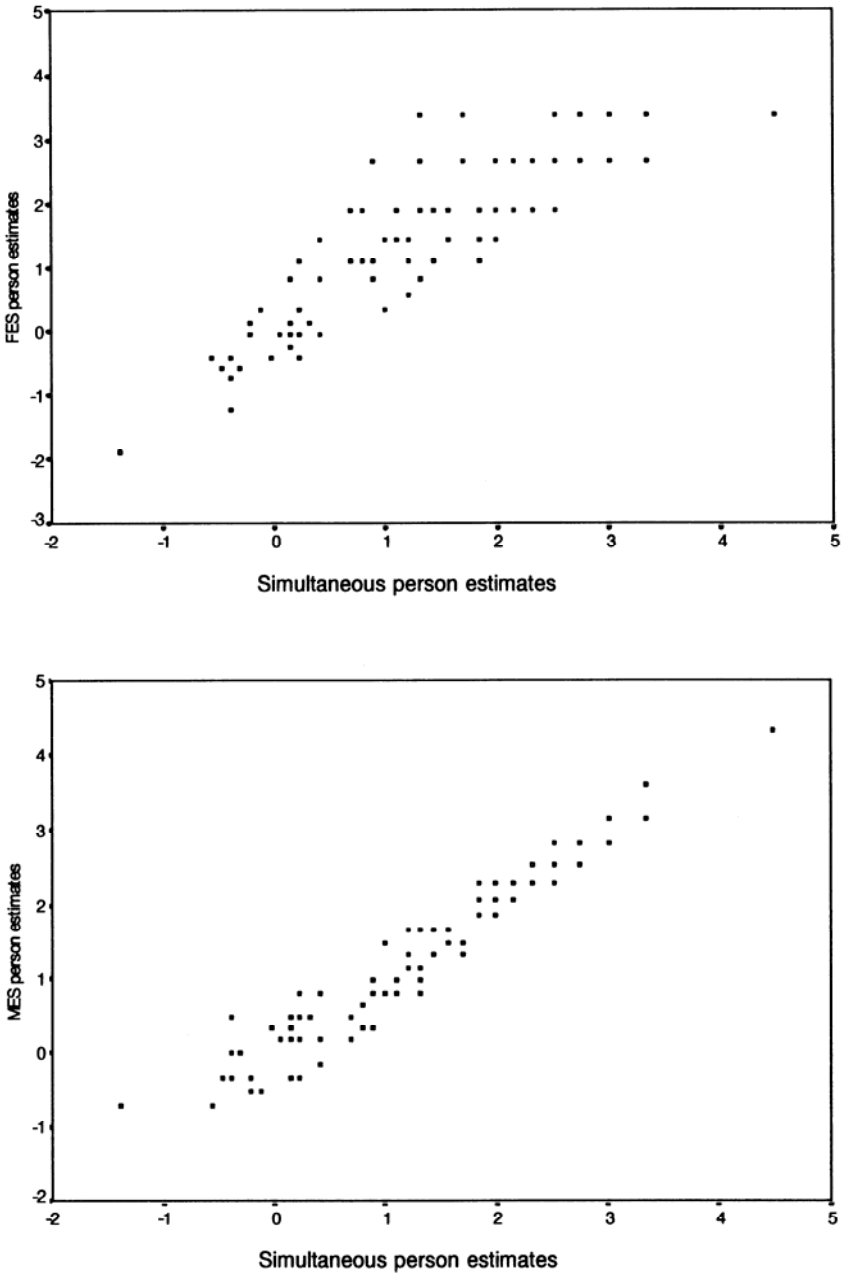
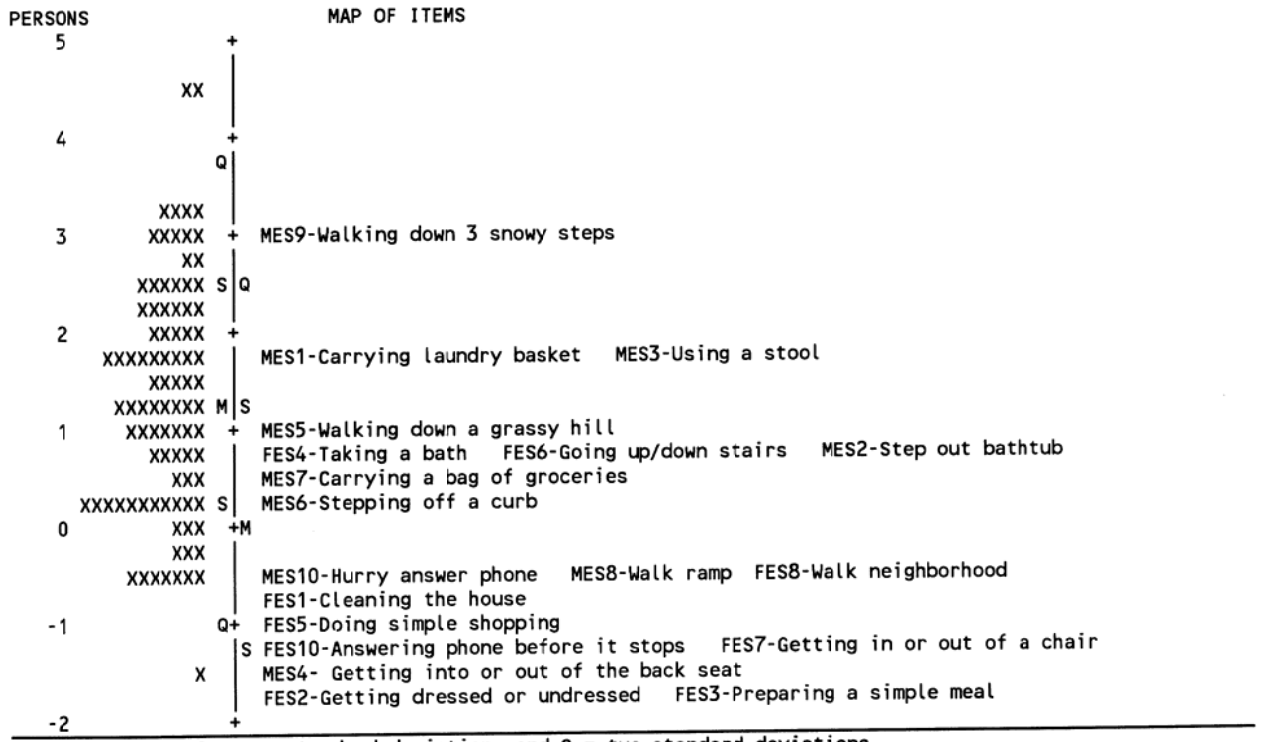


FIGURE 1 Correlation between person estimates form the simultaneous callibration and the anchored FES and MES items.



Note. M = mean, S = one standard deviation, and Q = two standard deviations.

FIGURE 2 Variable map displaying logit ability of persons and calibrations of the FES and MES items.

sons with higher self-efficacy and items more difficult to endorse are located toward the top of Figure 2 with persons of lower self-efficacy and items easier to endorse near the bottom. From this variable map a logical sequencing of the behaviors can be observed, supporting the construct validity of the obtained responses. Routine daily activities such as "Getting dressed or undressed" , "Preparing a simple meal", and "Answering the telephone before it stops ringing" are easy for this group of subjects to endorse. More risky tasks such as "Hurrying into another room to answer the phone", "Stepping off a curb without any help", and "Walking down a grassy hill" are found in the middle of the self-efficacy continuum. Finally, the most dangerous task, "Walking down 3 snowy steps without a handrail", is the most difficult activity to endorse for this group.

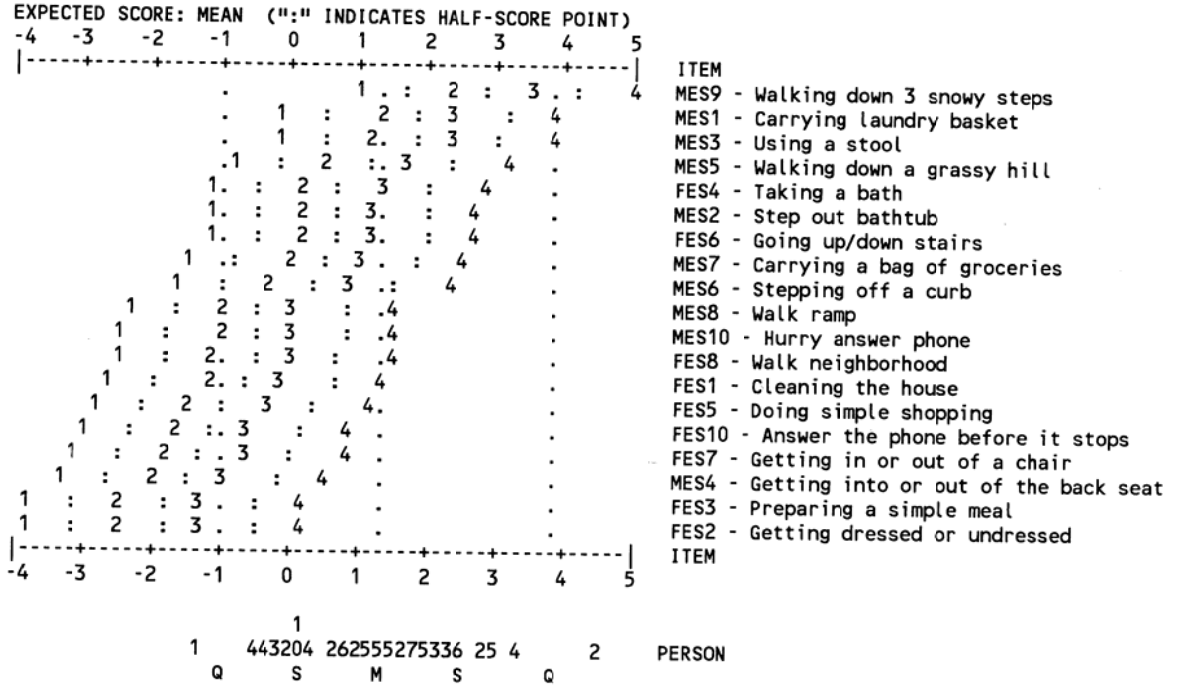
DISCUSSION

Attention to concepts such as self-efficacy in rehabilitation interventions for older adults with mobility impairment has the potential to positively impact outcome effectiveness. The social cognitive model identifies four sources of self-efficacy perceptions (Bandura, 1986, 1987, 1989). Successful performance has the most powerful and important influence on self-efficacy beliefs. Self-efficacy can also be augmented by vicarious learning from the observation of successful performances by a meaningful model or peer group, by verbal persuasion and feedback focusing on true ability and successful performance, and by diminishing physiological cues associated with stress while enhancing those associated with effective completion of the activity, task, or behavior. Practical strategies based on these four influences on self-efficacy can be integrated into goal setting and treatment planning. Attention to patient's current physical and psychological status, as well as the conditions of the environment provide a framework for goal setting and intervention activities at a level of difficulty which is sufficiently challenging to be interesting yet provides opportunities for success. Interaction with others who have faced similar challenges provides encouragement as well as alternative models and strategies for reaching rehabilitation goals. Reframing patients perceptions of lack of ability to one of lack of effort to meet task demands focuses on modifiable behaviors rather than underlying personal characteristics. Recognition of the adverse influences of anxiety and attention to reducing associated negative physiological consequences will assist realistic appraisals of ability and en-

hance performance of targeted activities.

Self-efficacy measures, like the FES and MES, used in conjunction with impairment level and functional assessment measures may assist rehabilitation professionals in setting goals, planning interventions, and assessing outcome effectiveness. Although the data presented here do not allow extrapolation to patients enrolled in treatment programs, a heuristic example of how the results obtained from the methodology presented may be used in planning a hypothetical treatment program should provide guidelines for future research involving older adults participating in treatment programs. For example, in planning a treatment program the behaviors, or proxies for the behaviors, since "Walking down 3 snowy steps without a handrail" is difficult to reproduce in a clinic, shown in Figure 2 which are easy to endorse indicate types of behavior for which treatment may be redundant. Difficult items indicate behaviors on which treatment should concentrate. "Getting dressing or undressed" is easy for all participating subjects. This indicates subjects are efficacious in attempting this behavior and treatment time may be better spent concentrating on more challenging behavior such as "Carrying a full laundry basket down the stairs". The variable map in Figure 2 recommends guidelines for planning group treatment. In order to maximize performance and capitalize on subjects self-efficacy appraisals, treatment time and the order in which behaviors are introduced should proceed from easiest too most difficult.

If tailored treatment plans are possible then a more detailed construct representation than provided in Figure 2 is required. The item calibrations in Figure 2 represent the overall difficulty of the behavior. One can place subjects relative to the behaviors, but what is not known is the subject's expected response on each of these behaviors as represented by the subject's location on the self-efficacy continuum. The expected score for each item can be obtained for each self-efficacy level by summing products of the probability of responding in a given category and the score for the same category. This information is depicted in Figure 3. Items are listed in order of decreasing difficulty and the distribution of self-efficacy estimates (i.e. the number of people scoring at each logit) is located below the horizontal axis. Take the three people at logit 2.15. This is represented by the darkened vertical line in Figure 3. Each of these subjects should not worry about spending a lot of energy and time on the behaviors for which a high level of self-efficacy is present (i.e. those behaviors which have an expected score of "4" which corresponds to "not at all concerned"). Behaviors



Note. M = mean, S = one standard deviation, Q = two standard deviations.

FIGURE 3 Expected response to each of the FES and MES items.

such as "Taking a bath or shower" and easier represent activities that these subject's demonstrate a high level of self-efficacy for accomplishing. A treatment program addressing these types of behaviors would be redundant, cost ineffective, and not motivating for those involved. Treatment time would be better spent on behaviors that were more difficult to endorse. Specifically, following goal setting strategies based on Social Cognitive Theory, activities with expected response scores (in this case integer values provide the most meaning as these are labeled positions on the rating scale) slightly above the subject's self-efficacy level should be targeted first (i.e. MES5, MES3, and MES1). Once these goals have been achieved, the process is repeated with the next set of activities with expected scores slightly above the subject's new self-efficacy level. This detailed type of treatment plan reduces treatment costs and increases the likelihood that subject's will not find the treatment plan redundant at their current level of perceived self-efficacy.

CONCLUSION

This article had three goals. First was to demonstrate the FES failed to assess the higher ends of the self-efficacy continuum. Second was to address this lack of scale definition by examining the psychometric properties of the simultaneous calibration of two assessments designed to measure self-efficacy for avoiding falls in elderly persons. Third, this article demonstrates the utility of using Rasch analyses of self-efficacy responses to plan a theoretically supported treatment plan.

The results demonstrated the inadequacy of the FES to assess the higher ends of the self-efficacy continuum. With the inclusion of the MES items, better scale definition was obtained. It is therefore recommended that the FES not be used in isolation to obtain estimates of self-efficacy. Either simultaneous use of both assessments or, if forced to choose, the MES provides a more precise estimate of self-efficacy perceptions than the FES.[1]

Modifications of the simultaneously scaled items should also be considered in future applications. With an average person measure of 1.35 logits, the simultaneously calibrated items demonstrated a better match between item difficulty and person self-efficacy estimates than the calibration of the FES items alone, which produced an average person measure of

1.89 logits. However, despite this improvement, the simultaneously calibrated items still did not provide an optimal match between item difficulty and person self-efficacy estimates. Specifically, scale definition may be improved by creating new items designed to fill in gaps found in the current calibration. This is particularly relevant at the higher end of the self-efficacy continuum where large areas above one standard deviation in Figure 2 demonstrated a lack of item sampling. In addition, plots of standardized INFIT and OUTFIT values against item calibrations revealed that the FES item "Walking around the neighborhood" (INFIT=-2.1, OUTFIT=-2.1) was not contributing much to the scale definition shared by the remaining items and may need modification or removal. Redundancy among the items referring to walking in various situations may be contributing to the negative misfit. Of course, given a Type I error rate of 5% in detecting misfitting items, one item was expected to be identified as not fitting the Rasch model. Replication of the current findings should help to clarify this situation.

In summary, self-efficacy beliefs are important factors to consider when attempting to engage subjects in a particular set of behaviors. Subjects who believe that they have the skills for a given behavior are more likely to attempt the behavior in the future. Findings from this initial work demonstrates a valuable tool for assessing perceived self-efficacy for the purpose of planning treatment. The measurement of self-efficacy may also prove to be a cost effective method of gauging ability as this type of self-report data is cheaper and faster to administer than collecting performance measures.

REFERENCES

Baker, S.P., & Harvey, A.H. (1985). Fall injuries in the elderly. *Clinics in Geriatric Medicine, 1*, 501-512.

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist, 44*, 1175-1184.

Bandura, A. (1987). Reflections on self-efficacy. In S. Rachman (Ed.), *Advances in behavior research and therapy.* Oxford: Pergamon Press.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Bhala, R.P., O'Donnell, J., Thoppil, E. (1982). Ptophobia: Fear of falling and its clinical management, *Physical Therapy, 62*, 187-190.

Borg, W.R., & Gall, M.D. (1989). *Educational research* (5th ed.). New York: Longman.

Buescher, K.L., Johnson, J.A., Paker, J.C., Smarr, K.L., Buckelew, S.P., Anderson, S.K., & Walker, S.E. (1991). Relationship of self-efficacy to pain behavior. *Journal of*

Rhuematology, 18, 968-972.

Campbell, A.J., Borrie, M.J., & Spears, G.F. (1989). Risk factors for falls in a community based prospective study of people 70 years and older. *Journal of Gerontology, 44*, M111-117.

Campbell, A.J., Reinken, J., Allan, B.C., & Martinez, G.S. (1981). Falls in old age: A study of frequency and related clinical factors. *Age and Aging, 10*, 264-270.

Chandler, J.M., & Duncan, P.W. (1993). Balance and falls in the elderly: Issues in evaluation and treatment. In A.A. Guccione (Ed.), *Geriatric Rehabilitation*. St. Louis: Mosby.

DeVito, C.A., Lambert, D.A., Sattin, R.W., Bacchelli, S., Ros, A., Rodriques, J.G. (1988). Fall injuries in the elderly: Community based surveillance. *Journal of the American Geriatrics Society, 36*, 1029-1035.

Downton, J.H., & Andrews, K. (1990). Postural disturbance and psychological symptoms amongst elderly people living at home. *International Journal of Geriatric Psychiatry, 5*, 93-98.

Evans, J.G. (1988). Falls and fractures. *Age and Aging, 17*, 361-364.

Fiatarone, M.A., Marks, E.C., Ryan, N.D., Meridith, C.N., Lizsitz, L.A., & Evans, W.J. (1990). High intensity strength training in nonagenarians: Effects on skeletal muscle. *Journal of the American Medical Association, 263*, 3029-3034.

Froman, R.D., & Owen, S.V. (1991). High school students' perceived self-efficacy in physical and mental health. *Journal of Adolescent Research, 6*, 181-196.

Gable, R.K., & Wolf, M.B. (1993). *Instrument development in the affective domain* (2nd ed.). Boston: Kluver Academic Publishers.

Hindmarsh, J.J., & Estes, E.H. (1989). Falls in older people, causes and intervention. *Archives of Internal Medicine, 149*, 2217-2222.

Holiday, P.J. (1992, June). Sequelae of falls: The fear of falling syndrome. Paper presented at the American Physical Therapy Association Annual Conference, Denver, CO.

Jette, A.M., Branch, L.G., & Berlin, J. (1990). Musculoskeletal impairments and physical disablement among the aged. *Journal of Gerontology, 46*, M203-208.

Kaplan, R.M., Ries, A.L., Prewitt, L.M., & Eakin, E. (1994). Self-efficacy expectations predict survival for participants with chronic obstructive pulmonary disease. *Health Psychology, 13*, 366-368.

Kilgore, K.M., Fisher, Jr., W.P., Silverstein, B., Harley, J.P., & Harvey, R.F. (1993). Application of Rasch analysis to the patient evaluation and conference system. In C. Granger & G. Gresham (Eds.), *Physical Medicine and Rehabilitation Clinics of North America: New Developments in Functional Assessment*, 4(3), pp. 493-515. Philadelphia: W.B. Saunders.

Linacre, J.M., & Wright, B.D. (1995). BIGSTEPS computer program. Chicago: MESA Press.

Lorig, K., & Holman, H. (1993). Arthritis self management studies: A twelve year review. *Health Education Quarterly, 20*, 17-28.

Magaziner, J., Cadigan, D.A., Hebel, J.R., & Parry, R.E. (1988). Health and living ar-

rangements among older women: Does living alone increase risk of illness? *Journal of Gerontology, 43,* M127-133.

Maki, B.E., Holiday, P.J., & Topper, A.K. (1991). Fear of falling and postural performance in the elderly. *Journal of Gerontology, 46,* M123-131.

Manton, K.G. (1988). A longitudinal study of functional change and mortality in the United States. *Journal of Gerontology, 46,* M123-131.

McAuley, R., Lox, C., & Duncan, T.E. (1993). Long term maintenance of exercise, self-efficacy, and physiological change in older adults. *Journal of Gerontology, 48,* P218-224.

Melton, L.J., & Riggs, B.L. (1985). Risk factors for injury after a fall. *Clinics in Geriatric Medicine, 1,* 525-539.

Mendes de Leon, C.F., Seaman, T.E., Baker, D.I., Richardson, E.D., & Tinetti, M.E. (1996). Self-efficacy, physical decline, and change in functioning in community living elders: A prospective study. *Journal of Gerontology, 51B,* S183-190.

Murphy, J., & Isaacs, B. (1982). The post fall syndrome: A study of 36 elderly patients. *Gerontology, 28,* 265-270.

Myers, A.M., Powell, L.E., Maki, B.E., Holiday, P.J., Brawley, L.R., & Sherk, W. (1996). Psychological indicators of balance confidence: Relationship to actual and perceived abilities. *Journal of Gerontology, 51A,* M37-43.

Nevitt, M.C., Cummings, S.R., Kidd, S. & Black, D. (1989). Risk factors for recurrent nonsyncopal falls: A prospective study. *Journal of Gerontology, 46,* M164-170.

O'Leary, A., Shoor, S., Lorig, K., Holman, H.R. (1988). A cognitive behavioral treatment for rheumatoid arthritis. *Health Psychology, 7,* 527-544.

Perry, B.C. (1982). Falls among the elderly living in high rise apartments. *Journal of Family Practice, 14,* 1069-1073.

Pfeiffer, E. (1975). A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. *Journal of the American Geriatrics Society, 23,* 433-441.

Powell, L.E., & Myers, A.M. (1995). The activities-specific balance confidence scale. *Journal of Gerontology, 50,* M28-34.

Prudham, D. & Evans, J.G. (1981). Factors associated with falls in the elderly: A community study. *Age and Aging, 10,* 141-146.

Reinsch, S., MacRae, P., Lachenbruch, P.A., & Tobis, J.S. (1992). Attempts to prevent falls and injury: A prospective community study. *Gerontologist, 32,* 450-456.

Rejeski, W.J., Craven, T. Wittinger, W.H., McFarlance, M, & Shumaker, S. (1996). Self-efficacy and pain in disability with osteoarthritis of the knee. *Journal of Gerontology, 51B,* P24-29.

Speechley, M., & Tinetti, M.E. (1991). Falls injuries in frail and vigorous community elderly persons. *Journal of the American Geriatrics Society, 39,* 46-52.

Tideiksaar, R., Silverton R. (1989). Psychological characteristics of older people who fall. *Clinical Gerontologist, 8,* 80-83.

Tinetti, M.E. (1992, June). Prevalence, morbidity, risk factors: A practical approach to evaluating and treating older persons at risk of falling. Paper presented at the American Physical Therapy Association Annual Conference, Denver, CO.

Tinetti, M.E., Mendes de Leon, C.F., Doucette, J.T., & Baker, D.I. (1994). Fear of falling and fall-related efficacy in relationship to functioning among community living elders. *Journal of Gerontology, 49*, M140-147.

Tinetti, M.E., Richman, D., & Powell, L. (1990). Falls efficacy as a measure of fear of falling. *Journal of Gerontology, 45*, P239-243.

Tinetti, M.E., Speechley, M., & Ginter, S. (1988). Risk factors for falls among elderly persons living in the community. *New England Journal of Medicine, 319*, 1701-1706.

Tinetti, M.E., Williams, T.F., & Mayewski, R. (1986). Fall risk index for elderly patients based on number of chronic disabilities. *American Journal of Medicine, 80*, 429-433.

Vellas, B., Cayla, R., Bocquet, H., dePemille, F., & Albarede, J.L. (1987). Prospective study of restriction of activity in old people after falls. *Age and Aging, 16*, 189-193.

Walker, J.E., & Howland, J. (1991). Falls and fear of falling among elderly persons living in the community: Occupational therapy interventions. *American Journal of Occupational Therapy, 45*, 119-122.

Wild, D., Nayak, U.S., Isaacs, B. (1981). How dangerous are falls in old people at home? *British Medical Journal, 24*, 266-268.

Wolf, S.L., Kutmer, N.G., Green, R.C., McNeely, E. (1993). The Atlanta FICSIT study: Two exercise interventions to reduce frailty in elders. *Journal of the American Geriatrics Society, 41*, 329-332.

Wolfson, L.I., Whipple, R., Judge, J., Amerman, P., Berby, C., & King, M. (1993). Training balance and strength in the elderly to improve function. *Journal of the American Geriatrics Society, 41*, 340-343.

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

Footnote

[1] An independent calibration of the MES items yielded the following information: Range of item calibrations 2.27 to -2.14; item reliability index .98; item challenge strata 10.87; person reliability index .84; person self-efficacy strata 3.41; and average person measure .68 (SD = 1.21). These indices demonstrate improvement in scale definition over the results presented for the independent FES item calibration.

JOURNAL OF OUTCOME MEASUREMENT, 1(1), 56-85

Dimensionality of an Early Childhood Scale Using Rasch Analysis and Confirmatory Factor Analysis

Madhabi Banerji
Pasco County School System, Florida

Richard M. Smith
Rehabilitation Foundation, Inc./Marianjoy Rehabilitation Hospital and Clinics

Robert F. Dedrick
University of South Florida

This paper explores the use of Rasch analysis and linear confirmatory factor analysis as methods for investigating the dimensionality of an early childhood test (Gesell School Readiness Screening Test), taking into account the theoretical basis of test construction. The paper presents the results of empirical analyses using both approaches and discusses the theoretical and psychometric considerations that guide the selection and application of each technique.

The first author is currently affiliated with the Department of Educational Measurement and Research, University of South Florida.
An earlier version of this paper was presented at the Annual Meeting of the Florida Educational Research Association, November, 1993.
Requests for reprints should be sent to Madhabi Banerji, University of South Florida, 4202 E. Fowler Ave., FAO-100U, Tampa, FL 33620

An important consideration in scale development is the dimensionality of data generated by an instrument. Dimensionality pertains to whether the instrument was constructed to represent a single attribute, defined by a coherent domain of behaviors, or of more than one logically distinguishable, subdomain of behaviors. In investigating dimensionality, we seek to answer the question: Can a unitary construct account for all the structure of the specified domain (Messick, 1989), thus justifying the use of a single, domain measure to describe the extent to which the variable exists in individuals? If the data are multidimensional, two sets of questions arise for measurement practitioners. First, how can multidimensionality be determined--i.e., what techniques and criteria will help to make useful judgments about multidimensionality of the data? Second, what implications does a multidimensional data structure have for measure interpretation and use-- i.e., should one make inferences from subdomain measures and ignore the overall measure, and, when and how is it reasonable to combine the subdomain measures despite the apparently separate data structures of the subdomains?

Such questions have not been satisfactorily addressed in the measurement literature, although they impinge directly upon construct validity. This paper explores the above issues from a measurement practitioner's point of view, by employing two analytic techniques, Rasch analysis and confirmatory factor analysis (CFA), to examine the structure of data generated from an early childhood test, the Gesell School Readiness Screening Test. The study uses an actual data set to examine the utility of each psychometric approach for detecting dimensionality, given that information is available on the theoretical basis underlying the test.

TRADITIONS OF SCALE DEVELOPMENT AND VALIDATION

The two psychometric approaches considered in this paper, Rasch analysis and confirmatory factor analysis, have their roots in different measurement traditions. Each psychometric tradition differs in terms of purposes of scaling, the mathematical and logical processes of scale construction (how the variable is operationalized), and the criteria used to evaluate dimensionality.

Rasch Psychometric Approach

The Rasch psychometric method (Andrich, 1988; Wright & Masters, 1982; Wright & Stone, 1979) is rooted in the tradition of Thurstone scaling, and locates the positions of items on an underlying continuum. The Thurstone tradition is described as the stimulus-centered approach (Crocker & Algina, 1986). The logical process of scale construction involves discrimination of the differences among stimuli (items) with respect to the amount of attribute present in each, using human judgment. Thurstone (1927) applied the method of *paired comparisons,* in which judges compared a set of attitude statements to one another, and used the measure of dispersion in the judgments to allocate item positions on a continuum. In another application, where Thurstone and Chave (1929) measured attitudes towards the church with the method of *equal-appearing intervals,* judges ordered a group of attitudinal items in 11 categories that they perceived to be equidistant with respect to their degree of favorability towards the church. Stimulus-centered scaling procedures specify methods of data collection, equations for estimating scale values, and statistical tests of goodness-of-fit between observed and estimated scale values (Crocker & Algina, 1986).

The Rasch measurement models are mathematical probability models that enable the examination of unidimensionality and ordering of items on a measurement continuum--fundamental requirements of Thurstone scaling (Andrich, 1988; Smith, 1992). When observations on a hypothetical construct are collected by empirical processes implying order (Wright & Masters, 1982), the Rasch models provide a way of transforming the ordinal observations into measurements that have the critical properties of linearity and specific objectivity. Items and persons are measured on a common interval scale. An additional, desirable consequence of employing Rasch models to define a measurement continuum, is that estimates of item measures and person measures are independent of one another (item-freed person measures, and person-freed item measures), thus making the measurements truly objective (Wright, 1967).

Criteria for dimensionality using Rasch analysis. In Rasch applications, the properties of unidimensionality and additivity depend on the extent to which the data gathered from the instrument fit the requirements of the Rasch model that is applied to transform the observations into measures. Assessment of fit at the item and person level is a common way by which dimensionality is examined by Rasch psychometricians (Smith, 1992;

Smith, 1996; Wright, 1996). Detecting dimensionality involves establishing invariance of estimated item and person measures in subsets of the data with the help of tests of fit. When an item misfits, it means that the item fails to discriminate between high and low performers in a way that is consistent with other items. When persons misfit, it means that their responses are inconsistent with the pattern of responses for people with similar ability measures.

Different Rasch models impose different expectations on the data generated by an instrument. The dichotomous model expects an ordering of 0,1 scored items in increasing degrees of difficulty, and unidimensionality of the data is established based on whether the obtained frequencies of correct responses for persons of varying ability are in agreement with the expected probabilities given by the model (Wright & Stone, 1979). The rating scale and partial credit models (Wright & Masters, 1982) expect data to be produced from multi-step tasks or items. Such items have to be polychotomously scored with a rating scale that implies an ordering of steps by difficulty within each item or task (Wright & Masters, 1982). Dimensionality in polychotomously scored data is tested by examining fit of the data with the expected probabilities given by the Rasch rating scale or partial credit models, depending on the particular application.

In examining dimensionality with Rasch techniques, then, it is important for practitioners to use careful judgment in selecting a Rasch model that is theoretically and logically consistent with the operationalization process of the construct.

Factor Analytic Approach

Factor analysis attempts to explain the covariation among a set of observed variables, the items, in terms of a set of underlying dimensions, the factors (Long, 1983). In linear factor analysis, each observed variable is conceptualized as a linear function of one or more factors. Factor analysis is commonly applied using exploratory or confirmatory approaches, with different mathematical procedures employed in each.

Exploratory factor analysis (EFA) examines how *all* the observed variables relate to *all* possible latent factors, typically with the help of principal components or principal axis factor extraction techniques. Confirmatory factor analysis (CFA) is used to test the fit of the data generated from items with a theoretically postulated factor structure, called a covariance structure model. By imposing theoretically motivated constraints on a struc-

tural model representing relationships among latent and observed variables (i.e., factors and items), CFA enables an evaluation of fit of the sample covariance matrix to an estimated population covariance matrix. Parameter estimates that reproduce the sample convariance matrix might be obtained using several estimation procedures including maximum likelihood, generalized least squares or weighted least squares (Hoyle, 1995).

Factor analysis is typically associated with subject-centered approaches to scale construction. Subject-centered approaches, sometimes referred to as the Likert tradition in measurement, aim to scale subjects rather than items, on a continuum (Crocker & Algina, 1986). In this approach, persons who take a test are placed on a continuum based on their level of performance on a domain or subdomain.

Domain sampling is a common method of scale development in the subject-centered approach (Crocker & Algina, 1986). Here, the items are sampled from a hypothetical, behavioral domain representing the construct. Items within a domain or subdomain operationally define the construct, but are not intentionally ordered by amount of attribute present (degree of difficulty). For all practical purposes, items are thought to be interchangeable with respect to their degree of difficulty or location on the continuum, as the subject-centered approach is not concerned with scaling of items. Persons who take the test, on the other hand, are placed on a continuum based on the magnitude of measures on items in a subdomain or domain.

Properties of a scale resulting from the application of the subject-centered approach to scale construction are typically examined by correlations. According to Messick (1989) "correlational evidence is (considered) highly relevant to appraising whether the degree of homogeneity in the test is commensurate with the degree of homogeneity expected from the construct theory of the domain" (p. 38). Item-total correlations, internal consistency estimates such as Cronbach's alpha and K-R 20, and factor analytic techniques are applied to investigate psychometric properties of scales derived using the subject-centered approach. Factor analysis, whether EFA or CFA, is commonly used for understanding dimensionality of data produced from test administrations.

Criteria for dimensionality using factor analysis. Criteria for determining the number of dimensions in a construct differ depending on whether EFA or CFA procedures are used. When EFA is used, the number of latent dimensions is typically identified by the researcher based on one or more of the following criteria: scree plot of eigenvalues; eigenvalues

greater than one; number of factors that are better than chance; magnitude of the loadings of items (regression coefficients) on factors; magnitude of the interfactor correlations; and extent to which loadings yield simple structure after rotation.

When CFA is used, the number of common factors and their interrelations are specified along with how each observed variable is related to each factor. Goodness-of-fit tests are used to evaluate the fit of the data to the specified dimensions in the model. To yield meaningful results, it is necessary that the structural model in CFA is consistent with the hypothesized domain structure of the construct being studied. (This principle applies to Rasch analysis applications as well).

Dimensionality is evaluated using model fit statistics and other supporting statistics such as standardized regression coefficients and R^2 values. Currently, there is no agreed-upon method for evaluating fit. Commonly used indices are the χ^2 likelihood ratio test, χ^2/df, Bentler's (1990, 1992) normed comparative fit index (CFI), and the root mean square error of approximation (RMSEA; Browne & Mels, 1990; Steiger & Lind, 1980).

An example of instrument development and validation that adheres to the subject-centered measurement tradition is found in the studies conducted by Marsh and colleagues in the development of a self-concept scale for preadolescents called the Self Description Questionnaire (see Marsh, 1987).

AN EMPIRICAL STUDY

Gesell School Readiness Screening Test

The Gesell School Readiness Screening Test (GSRT; Ilg, Ames, Haines, & Gillespie, 1978), the focus of the present empirical analysis, was designed to measure a general behavior domain, entitled Developmental Age (DA), in children from 2-8 years of age. The test is comprised of eight multistep tasks, each of which can be scored on a 14-point ordinal scale (0-13), representing increasing levels of DA from three years to seven years. According to the authors, DA is composed of two logically-derived subdomains-- Adaptive and Language behaviors. Five of the eight tasks of the GSRT are perceptual-motor tasks and fall under the Adaptive domain (Cubes or block-building, Copy Forms, Incomplete Man, Writing Name, and Writing Numbers). The remaining three tasks are verbal interviews, grouped by the authors under Language behaviors (Interview, Animals,

and Interests). A typical GSRT task, such as Copy Forms, requires the child to copy progressively complex geometric shapes, starting with a line and ending in a diamond. DA scores are assigned based upon the level of complexity of the shape that the child can successfully complete.

The GSRT tasks are scored using a DA score that takes values such as 3 years, 3 1/2 years, 3 1/2 - 4 years. To enable statistical treatment of the data, the DA values were rescaled using the conversion scheme in Table 1.

The authors recommend that an overall DA score be used to make inferences about a child's developmental maturity, but also recommend the subdomain scores. Conceptually, Adaptive and Language behaviors appear to be related under the broad domain of Developmental Age.

Both one- and two-factor solutions of the GSRT have been examined using EFA procedures (Banerji, 1992). The one-factor solution indicated that the eight tasks had loadings ranging from .57 to .83 on the factor, which explained 56% of the total variance among tasks. When a two-factor solution was pursued, correlated clusters of tasks were found that corresponded to the tasks grouped by the developers under the Language and Adaptive subdomains, explaining 60% of the total variance. The loadings of subdomain tasks on the separate factors ranged from .45 to .83, and the interfactor correlation was .72. Because of the high overlap of the factors, the EFA findings appeared to support the use of an overall DA score rather than separate subdomain scores. At the same time, the results suggested that the two factors provided some unique information.

One purpose of the present study was to evaluate the dimensionality of the GSRT data using Rasch and CFA procedures. A more important purpose, was to explore the utility of Rasch and CFA techniques, separately and in combination, for examining dimensionality of data from tests such as the GSRT. Both approaches could apply, as the GSRT was designed with a theoretical framework that could be tested with one- and two-factor CFA models; additionally, an ordering of difficulty was implied in the multi-step structure of each task, which could be investigated with Rasch analysis.

Table 1

Developmental Age Scoring Categories (Original and Collapsed)

	Original Scale													
Assigned DA	3.0	Between 3.0-3.5	3.5	Between 3.5-4.0	4.0	Between 4.0-4.5	4.5	Between 4.5-5.0	5.0	Between 5.0-5.5	5.5	Between 5.5-6.0	6.0	Between 6.0-7.0
Points	0	1	2	3	4	5	6	7	8	9	10	11	12	13
	Collapsed Scale													
Points	0	0	1	1	2	2	3	3	4	4	5	5	6	6

Note. DA = Developmental Age on Gesell School Readiness Screening Test.

METHOD

Data Source

The data for this study came from a stratified, random sample of kindergarten students (N = 523) from the 1988-89 population of kindergartners at a mid-size school system in central Florida. Elementary schools in the district served as the strata in the population. All children in the sample were administered the GSRT by trained examiners during the summer and early fall of 1988. Some cases had missing information on selected tasks of the GSRT, providing 509 cases for the CFA. The Rasch analysis was conducted using all 523 cases. The sample consisted of 250 males (51%) and 249 females (49%). The composition of the sample by ethnic category was 428 White (84%), 31 Black (6%), and 30 Hispanic (6%), 5 Asian/Pacific Islanders (1%), and 15 unknown (3%).

Analytic Procedures

Two approaches were taken in conducting the Rasch analysis of the GSRT data. First, since overall DA is determined based on all eight tasks, the fit of the total set of tasks to the model was tested using two response models and scoring scales. Second, the fit of the tasks to the logically and empirically derived subsets of tasks was checked. To test the fit of the tasks to the Rasch family of measurement models, the data were calibrated with the BIGSTEPS program (Wright & Linacre, 1996).

Rasch techniques provide two analytic options for examining dimensionality of polychotomous data, the rating scale model and the partial credit model. Both models assume that the items (tasks) are scored on a multi-point scale, as is the case with the GSRT. The rating scale model assumes that all tasks share the same step difficulties at successive points of the scale. The partial credit model allows the step difficulties to vary at successive points on the scale from task to task. Both rating scale and partial credit models were applied to verify whether the ordering of steps by difficulty was consistent across the eight tasks.

A series of analyses were conducted to determine whether the GSRT data fit the unidimensional structure given by the Rasch rating scale and partial credit models. Generally, the steps in the analysis were to: (1) calibrate the tasks using a selected model, (2) identify and

eliminate misfitting tasks, (3) identify and eliminate misfitting persons, and (4) recalibrate and check fit of data to the model. Adjustments were made in the procedures when the results in particular stages of the analysis showed anomalies.

First, the data were analyzed with the rating scale model using the 0-13 scoring scheme shown in Table 1. This analysis revealed that the scoring categories were not monotonically increasing, a result that was suspected to cause step misfit. The rating scale analysis was thus repeated after collapsing adjacent scoring categories to form a scale from 0 to 6, as shown in the bottom half of Table 1. As the misfit values were only marginally improved by adjusting the scoring scheme, it was decided that the rating scale model might be inadequate for these data. Subsequent analyses used the partial credit model and the collapsed (0 to 6) scoring scheme. Systematic elimination of misfitting tasks and persons resulted in a grouping of tasks that fit the unidimensional structure given by the Rasch partial credit model.

Confirmatory Factor Analysis Models

Two CFA models were examined that were congruent with the theoretical construct of Developmental Age underlying the GSRT: (a) a two-factor model in which five tasks (Cubes, Copying Forms, Incomplete Man, Writing Name, and Writing Numbers) loaded on the Adaptive factor and three tasks (Interview, Animals, and Interests) loaded on the Language factor (Language), and (b) a one-factor model in which the eight tasks loaded on one factor. Figure 1 displays the hypothesized two- and one-factor models, respectively. All CFA analyses used the 0 to 6 scoring scheme.

Each model was estimated using the weighted least squares (WLS) fitting function in LISREL 8 (Jöreskog & Sörbom, 1993). This method assumes that the observed variables are represented on an ordinal scale and analyzes the matrix of polychoric correlations with the corresponding asymptotic covariance matrix. The matrix of polychoric correlations and the asymptotic covariance matrix were obtained using PRELIS 2.

Given the well known limitations of the χ^2 and χ^2/df as measures of model fit (see Marsh, Balla, & McDonald, 1988 for a discussion of the effect of sample size on the χ^2), fit of the data to the CFA models was examined using statistics less sensitive to sample size. These fit statistics included the comparative fit index (CFI; Bentler, 1990) and

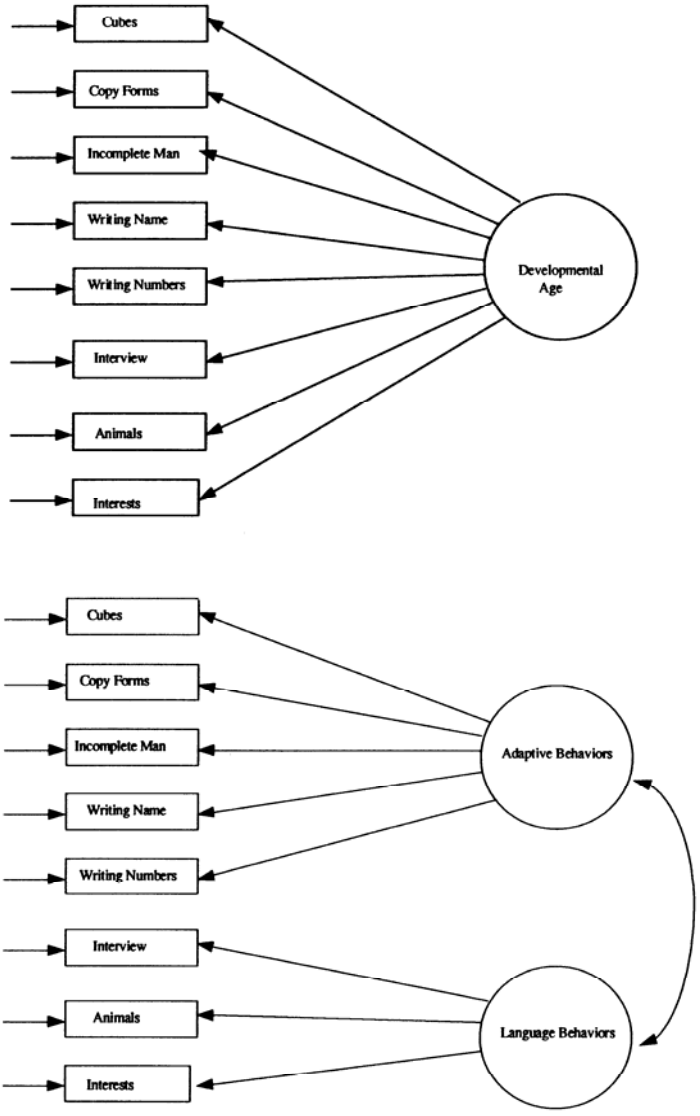


FIGURE 1 One- and two-factor confirmatory factor analysis models for the Gesell School Rediness Screening Test.

the root mean square error of approximation (RMSEA; Browne & Mels, 1990). Values greater than .90 on the CFI and values less than .08 on the RMSEA were viewed as indicators of acceptable fit (Rigdon, 1996). Multiple fit statistics were used because each has limitations, and there is no agreed upon method for evaluating whether the lack of fit of a model is substantively important (Marsh, Balla, & McDonald, 1988). The strategy used to evaluate overall model fit was to look for consistency across the indices and to consider "substantive, theoretical, and conceptual" factors (Jöreskog, 1971, p. 421) in addition to statistical criteria.

RESULTS

Rasch Analysis

The step calibrations for score categories following a rating scale analysis with the original (0-13) and revised (0-6) scheme are reported in Table 2. The findings were unsatisfactory on two accounts. First, there was considerable disorder in the step values, showing the odd-valued scoring categories to be greatly underutilized. The observed disorder might have resulted from the nature of the score reporting form, where examiners are asked to use the odd categories only when they were undecided. Second, there was a high degree of task or item misfit with six of the eight tasks having outfit values greater than +2.0, which is a standard criterion used in deciding the degree of item fit (Smith, 1991; Wright & Masters, 1982; Wright & Stone, 1979). The outfit mean was -1.0 (SD = 5.2); the expected mean and standard deviation of this statistic when the data fit the model are 0.0 and 1.0, respectively.

The results of the second rating scale analysis using the collapsed scoring category data indicated a considerable improvement in the ordering of the step difficulties, as shown in the bottom half of Table 2. However, some task misfit remained, with a mean outfit of -1.1 (SD = 5.0). Two of the eight tasks had an outfit value greater than a t value of +2.0.

Due to the continuing levels of misfit of two of the tasks, there appeared to be a strong possibility that the step difficulty structure was not the same for all eight tasks, and a partial credit analysis might be more useful for the data. The results of the first partial credit analysis of the collapsed category data indicated an improvement in the fit of the items to the model. The mean task outfit was -0.7 (SD = 4.0). How-

TABLE 2
Original and Revised Score Category Step Calibrations Rating Scale Model

Fourteen Score Categories

CATEGORY LABEL	STEP VALUE	OBSERVED COUNT	AVGE MEASURE	INFIT MNSQ*	OUTFIT MNSQ*	STEP MEASURE	STEP ERROR	EXPECTED SCORE MEASURES STEP-.5	AT STEP	STEP+.5	THURSTONE THRESHOLD	CATEG RESIDUL
0	0	22	-1.21	1.04	1.27	NONE			(-3.15)	-2.74		
1	1	3	-1.28	.43	.50	.86	.22	-2.74	-2.27	-1.94	-2.21	
2	2	183	-.91	1.07	1.16	-5.11	.21	-1.94	-1.67	-1.45	-2.17	-2.4
3	3	29	-.97	.47	.40	.98	.08	-1.45	-1.27	-1.12	-1.21	
4	4	291	-.66	.90	.89	-3.02	.08	-1.12	-.99	-.88	-1.15	-1.7
5	5	17	-.72	.11	.07	2.29	.06	-.88	-.76	-.65	-.77	
6	6	521	-.31	.89	.86	-3.78	.06	-.65	-.53	-.40	-.76	.7
7	7	240	-.13	.75	.70	.63	.05	-.40	-.25	-.08	-.39	.9
8	8	1157	.26	.89	.96	-1.48	.04	-.08	.11	.34	-.24	7.4
9	9	249	.53	.77	.93	1.88	.04	.34	.57	.83	.48	1.5
10	10	1014	.76	.94	.99	-.81	.04	.83	1.10	1.39	.64	2.0
11	11	140	.99	.98	.99	2.83	.06	1.39	1.73	2.16	1.55	-1.0
12	12	282	1.10	1.24	1.18	.39	.07	2.16	2.90	4.40	1.78	-6.2
13	13	14	1.33	1.30	1.21	4.33	.27	4.40	(5.44)		4.35	-.6
						modal			mean		median	

INFIT & OUTFIT ARE "OBSERVED MNSQ / EXPECTED MNSQ"
* Expected Value of MNSQ Ratio is 1.0

TABLE 2 Continued
Original and Revised Score Category Step Calibrations Rating Scale Model

Seven Score Categories

CATEGORY LABEL	STEP VALUE	OBSERVED COUNT	AVGE MEASURE	INFIT MNSQ	OUTFIT MNSQ	STEP MEASURE	STEP ERROR	EXPECTED SCORE MEASURES STEP-.5	AT STEP	STEP+.5	THURSTONE THRESHOLD	CATEG RESIDUL
0	0	25	-1.76	.99	1.16	NONE			(-4.71)	-3.78		-1.0
1	1	212	-1.16	1.01	1.07	-3.54	.21	-3.78	-2.58	-1.82	-3.64	-4.4
2	2	308	-.64	.82	.80	-1.22	.08	-1.82	-1.27	-.80	-1.61	-1.4
3	3	761	.14	.87	.85	-1.07	.06	-.80	-.30	.29	-.86	4.5
4	4	1406	1.24	.88	.95	.07	.04	.29	1.03	1.91	.20	9.9
5	5	1154	2.18	.94	.97	1.83	.04	1.91	2.94	4.19	1.87	-2.0
6	6	296	2.81	1.21	1.15	3.93	.07	4.19	(5.11)		4.03	-5.6
						modal			mean		median	

INFIT & OUTFIT ARE "OBSERVED MNSQ / EXPECTED MNSQ"

ever, two of the eight tasks had outfit values greater than +2.0. On the other hand, the fit of the persons was good (mean person outfit was -0.3, *SD* = 1.1), with only 18 of the 523 (3%) persons with an outfit value greater than +2.0. This rate is extremely close to the Type I error rate for this statistic (Smith, 1991).

Partial credit response category curves for each of the eight tasks are shown in Figure 2, and illustrate why the rating scale assumption of shared step values across tasks does not work for the GSRT. Had the data fit the rating scale model, the response category curves for all tasks would have been similar. As is clear, the response probability curves for task 2 (Copy Forms) and task 7 (Animals) are somewhat similar. The probability curves for task 4 (Write Name) and task 5 (Write Numbers), on the other hand, are quite different in terms of both height and shape.

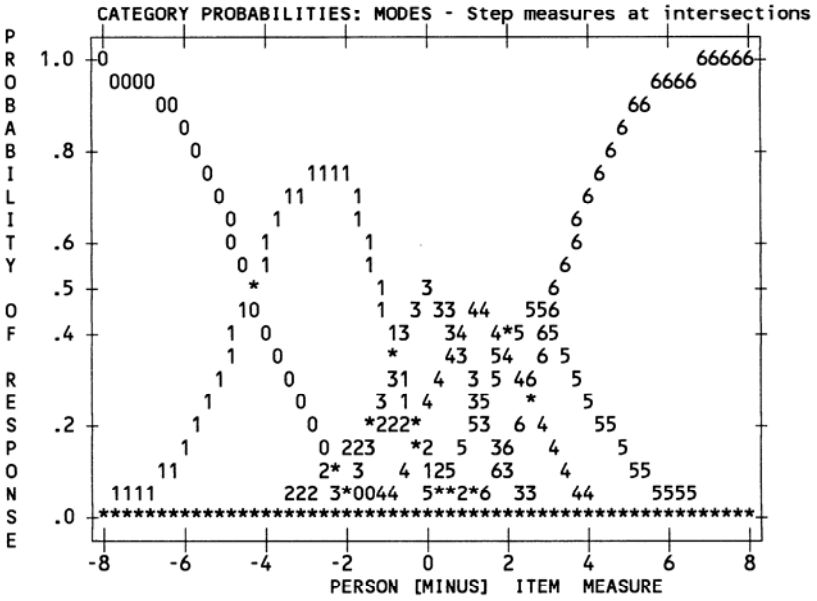
To be certain that misfitting persons were not causing the task misfit, all persons with outfit values equal to or greater than 1.6 were eliminated from the calibration of the total task set. In all 26 persons were eliminated and a slight improvement in task fit statistics was found. The mean item outfit statistic was -0.8 (*SD* = 3.2). Two out of the eight tasks on the scale had outfit values greater than +2.0 and suggested that the observed task misfit was due to causes other than person response patterns.

A summary of the results of the Rasch partial credit analysis, showing item difficulty calibrations, item-total point biserial correlations, and fit values of all eight tasks are shown in Table 3. The two misfitting tasks (1 and 7) had point biserial correlations of .58 (Cubes) and .61 (Animals), respectively.

The results of the next two partial credit analyses with collapsed scores focused on tasks divided into the two logical sets originally developed for the GSRT, Adaptive (tasks 1-5) and Language (tasks 6-8). The results of the Adaptive subset of tasks indicated a mean item outfit statistic of -1.3 (*SD* = 3.2). Task 1 (Cubes) had an outfit value greater than +2.0. The Language subset had a mean item outfit statistic of -0.5 (*SD* = 0.6). None of the three items had an outfit value greater than +2.0. These results suggested that, with the exception of task 1 (Cubes), the tasks fit the Rasch partial credit model when separated into their respective, logical subsets, and that individual tasks were causing the misfit rather than differences in the Adaptive and Language dimensions.

In the final analysis, the partial credit analysis was repeated with specific misfitting tasks removed. A plot of the task misfit found on the full range of eight tasks (i.e., prior to removing tasks with extreme fit

Item 1



Item 2

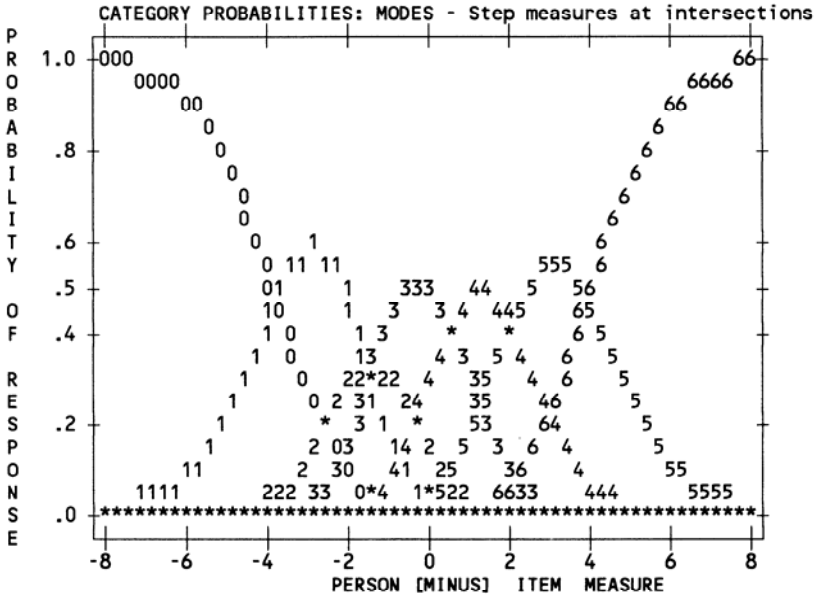
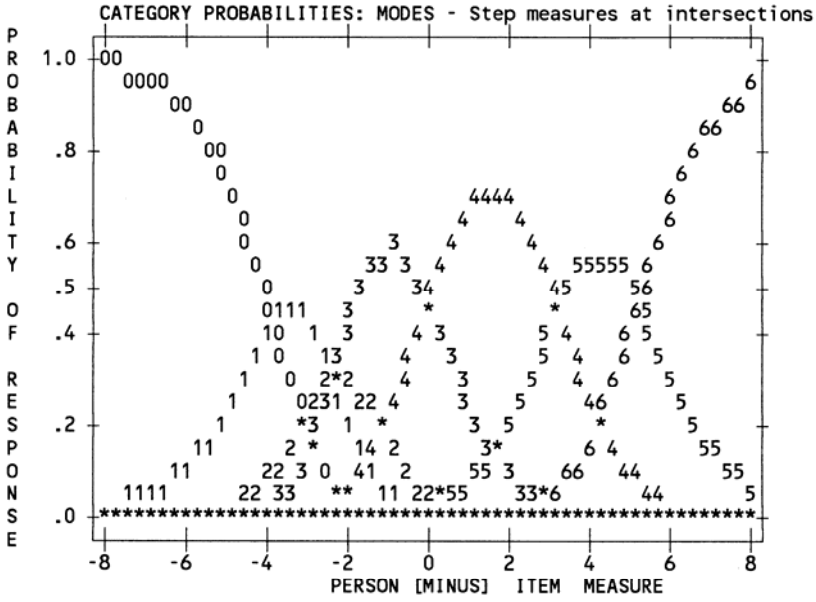


FIGURE 2 Response Category Probability Curves Partial Credit Model.

Item 3



Item 4

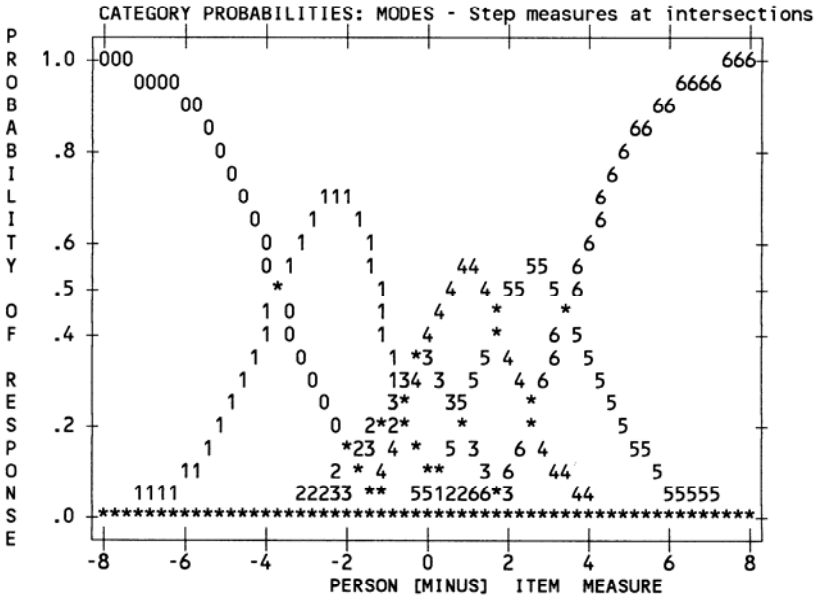
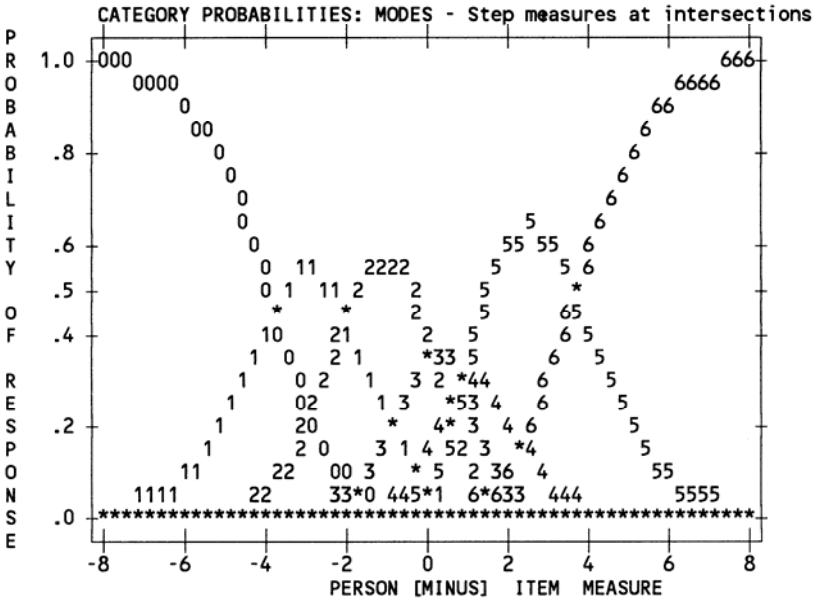


FIGURE 2 Cont'd Response Category Probability Curves Partial Credit Model.

Item 5



Item 6

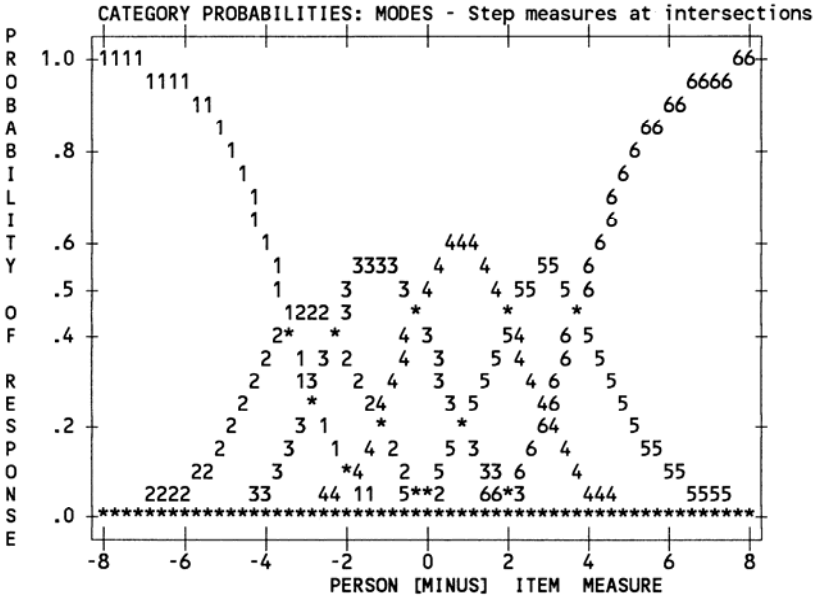
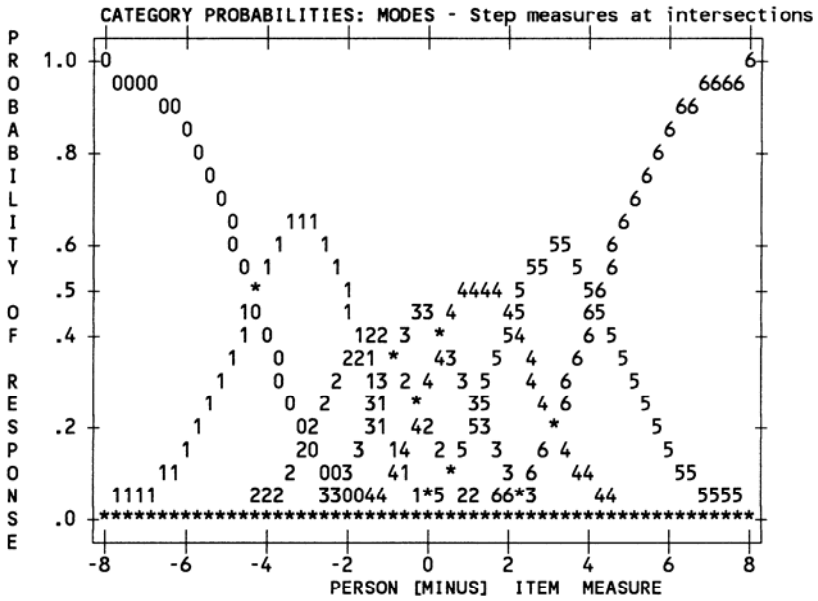


FIGURE 2 Cont'd Response Category Probability Curves Partial Credit Model.

Item 7



Item 8

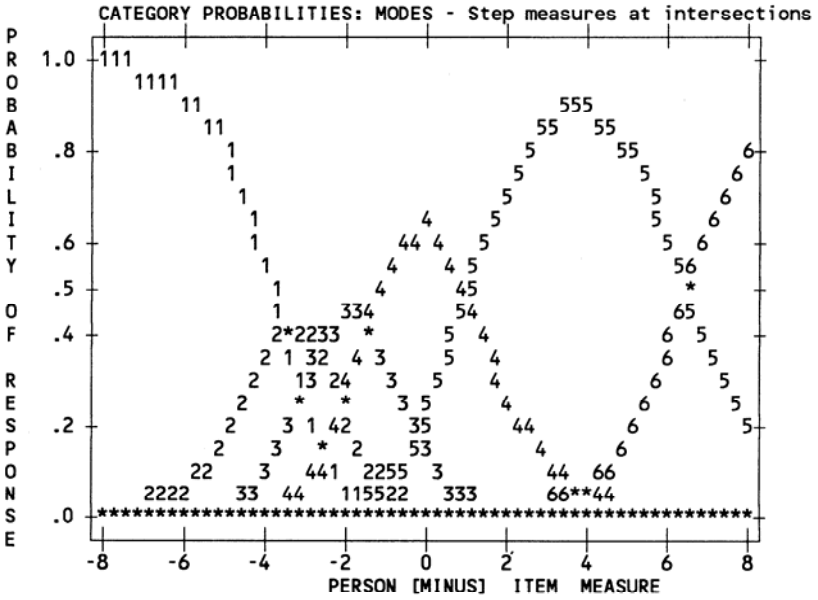


FIGURE 2 Cont'd Response Category Probability Curves Partial Credit Model.

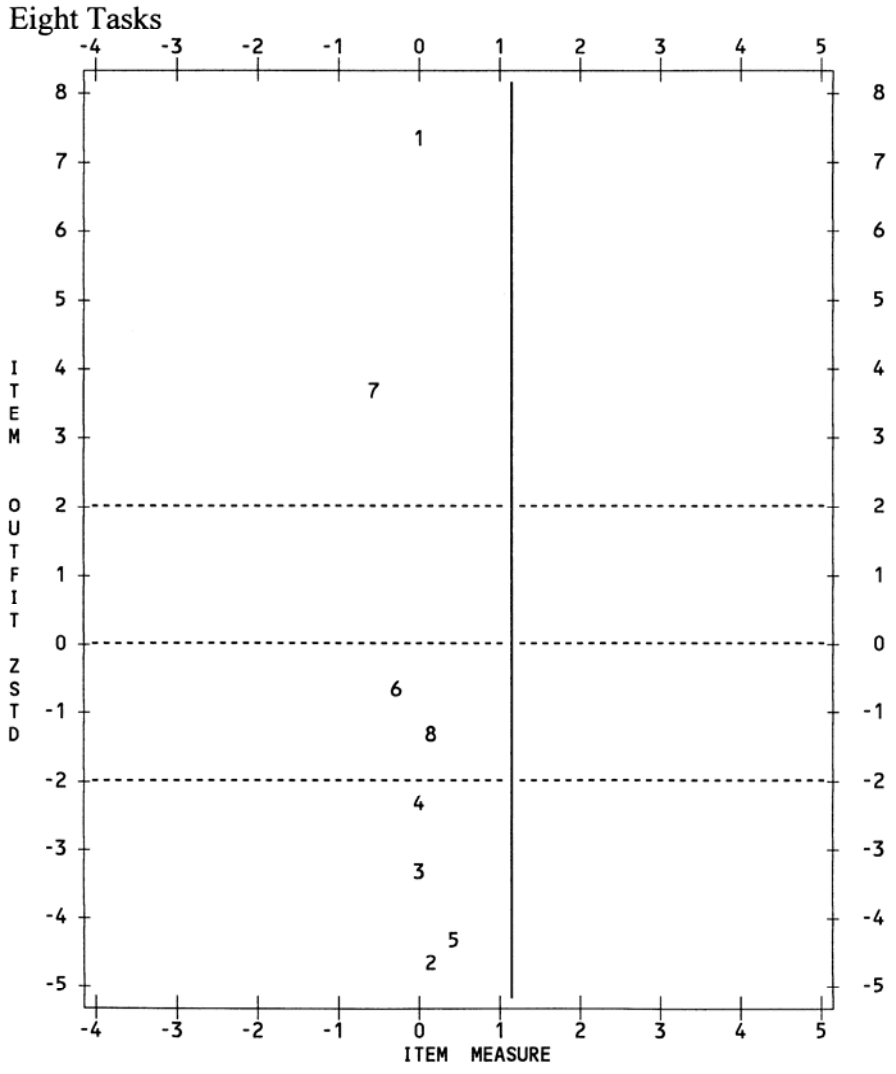


FIGURE 3 Fit plots for Eight and Six Tasks of the GSRT.

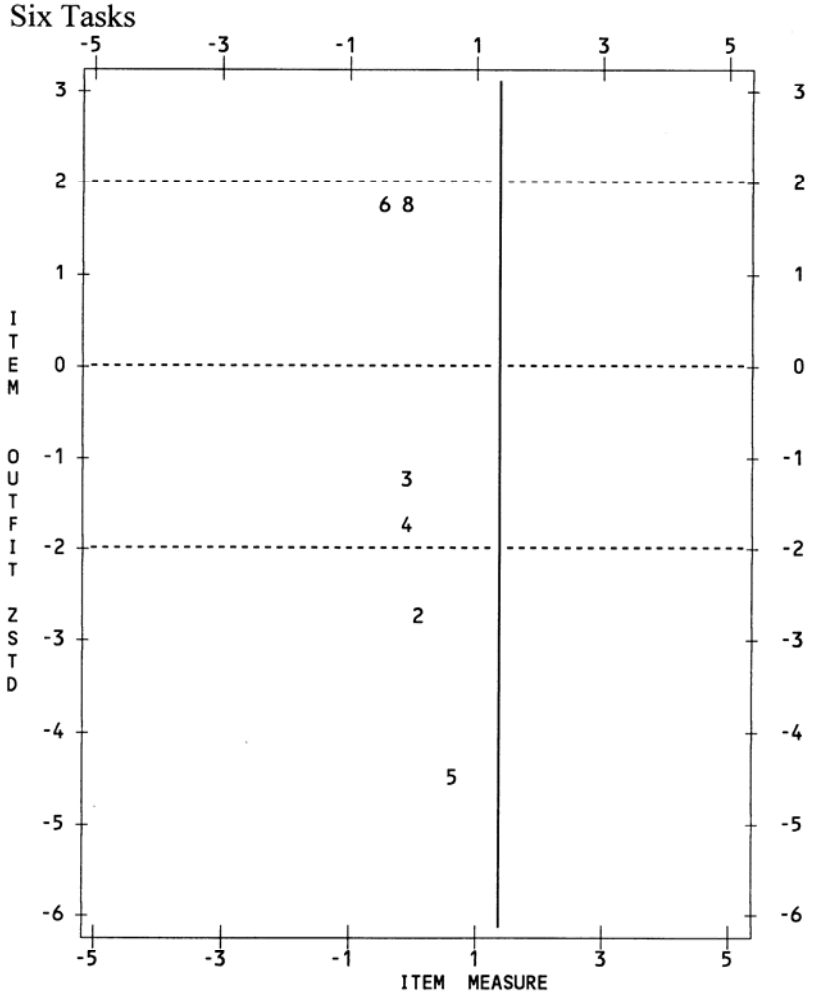


FIGURE 3 Cont'd Fit plots for Eight and Six Tasks of the GSRT.

TABLE 3
Summary Item Calibration Information Partial Credit Model

Seven Score Categories

ENTRY NUM	RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS CORR.	ITEMS	G
1	1954	522	.03	.05	1.52	7.3	1.54	7.4	.58	CUBES	0
2	1944	522	.13	.05	.74	-4.7	.74	-4.8	.79	COPY FORMS	0
3	1973	523	.02	.07	.81	-2.9	.79	-3.3	.72	INC MAN	0
4	2032	516	.00	.05	.85	-2.4	.85	-2.3	.76	WRIT NAME	0
5	1785	515	.49	.05	.74	-4.6	.75	-4.2	.77	WRIT NUMB	0
6	2209	522	-.28	.06	.97	-.4	.97	-.5	.66	INTERVIEW	0
7	2149	521	-.57	.06	1.23	3.5	1.25	3.8	.61	ANIMALS	0
8	2235	521	.18	.07	.91	-1.3	.91	-1.4	.66	INTRESTS	0
MEAN	2035.	520.	.00	.06	.97	-.7	.97	-.7			
S.D.	143.	3.	.29	.01	.25	3.9	.26	4.0			

TABLE 4

Real Person Separation Reliability Results Across Different Analyses

Analysis	Model[a]	Number of Score Categ.	Misfitting Persons	Number of Items	Person Sep. Index	Person Sep. Reliability
1	RS	14	Including	8	2.59	.87
2	RS	7	Including	8	2.58	.87
3	PC	7	Including	8	2.77	.88
4	PC	7	Deleted	8	2.99	.90
5	PC	7	Deleted	6	2.96	.90
6	PC	7	Deleted	5 (Adap)	2.73	.88
7	PC	7	Deleted	3 (Lang)	1.88	.78

[a] RS = Rating Scale
PC= Partial Credit

values) is shown in the first half of Figure 3. In this case, there are two clearly positive misfitting tasks, task 1 (Cubes) and task 7 (Animals). The four negative misfitting tasks tended to be less extreme and clustered, suggesting that if tasks 1 and 7 were removed, fit might improve. Deleting these two tasks from the analysis indicated a considerable improvement in the fit of the data to the unidimensional, partial credit model (see Figure 3). The mean item outfit for the six remaining tasks was -1.2 ($SD = 2.3$). None of the tasks had fit values greater than +2.0 as shown in the second half of Figure 3. The six best-fitting tasks consisted of tasks 2 (Copy Forms), 3 (Incomplete Man), 4 (Writing Name), 5 (Writing Numbers), 6 (Interview), and 8 (Interests).

The Rasch analytic methods yield information on reliability of a test through the person separation index and test reliability of person separation. The person separation index is a ratio of the unbiased standard deviation of the persons' logit measures to the standard error of measurement. It is interpreted to be conceptually equivalent to coefficient alpha. The person separation reliability of the GSRT, with tasks 1 (Cubes) and 7 (Animals) removed, showed no change following elimination of those items from the scale (See Table 4). This value was .90 for all eight tasks and .90 for the set of six tasks. The person separation reliability values for the Adaptive and Language task clusters were .88 and .78, respectively.

The removal of misfitting tasks showed slight improvement in the fit of the person responses to the two models. In the original eight task test there were 26 persons with outfit values greater than +1.5. In the six task test (tasks 1 and 7 removed) there were 25 misfitting persons, of which six were positive.

The results of the Rasch analysis suggest that there is a single underlying variable that is best described by six tasks: task 2, Copy Forms; task 3, Incomplete Man; task 4, Writing Name; task 5, Writing Numbers; task 6, Interview; and task 8, Interests. Despite the fact that these tasks come from two different theoretical dimensions, they function as a unidimensional unit. The deletion of the two misfitting tasks (tasks 1 and 7) improved the fit of the remaining tasks and persons to the Rasch partial credit model while losing no test reliability.

Confirmatory Factor Analysis

The χ^2 was statistically significant for the two-factor model using all eight GSRT tasks [$\chi^2(19, N = 509) = 66.39$, $p < .001$; $\chi^2/\text{df} = 3.49$].

Using the χ^2 as a fit statistic is problematic, however, as it is strongly influenced by sample size, and thus even small differences between the hypothesized model and observed data will result in statistically significant χ^2 values. When alternative measures of fit, less sensitive to sample size were used, the results indicated that the fit of the two-factor model was acceptable (CFI = .98 and RMSEA = .07) and was better than the fit for the one-factor model [$\chi^2(20, N = 509) = 115.79, p < .001$; $\chi^2/df = 5.79$; CFI = .96]. The RMSEA of .10 for the one-factor model indicated a less than acceptable fit.

Table 5 presents the weighted least squares estimates for the two-factor model. All of the parameter estimates were statistically significant ($p < .05$). The lowest factor loading within the Adaptive domain was .67 for Cubes, while the lowest factor loading within the Language domain was .73 for Animals. The correlation between the Adaptive and Language factors was .85 ($SE = .02$). The 95% confidence interval for this correlation (.81 to .89) indicated that there was considerable overlap between these factors, but there was some unique variance not shared by the factors. In the one-factor model (see Table 5), Cubes and Animals had the lowest factor loadings on the overall factor of Developmental Age (.66 and .69, respectively).

As a follow-up to the findings of the Rasch analysis, which indicated a grouping of six tasks (Copy Forms, Incomplete Man, Writing Name, Writing Numbers, Interview, and Interests) in a single, well-defined variable rather than the original set of eight tasks of the GSRT, two additional CFA models were tested. The first examined a two-factor model consisting of four tasks from the Adaptive domain (Copy Forms, Incomplete Man, Writing Name, and Writing Numbers) and two tasks from the Language domain (Interview and Interests); the second tested a one-factor model defined by the six tasks identified as unidimensional by the Rasch approach.

Results for the six tasks of the GSRT paralleled those of the eight tasks with the two-factor model providing better fit [$\chi^2(8, N = 509) = 27.45, p < .001$; $\chi^2/df = 3.43$; CFI = .99] than the one-factor model [$\chi^2(9, N = 509) = 56.08, p < .001$; $\chi^2/df = 6.23$; CFI = .98]. The RMSEA of .07 for the two-factor model indicated an acceptable fit, while the RMSEA of .10 for the one-factor model indicated less than acceptable fit. In the two-factor model, loadings within the Adaptive domain ranged from .78 (Incomplete Man) to .91 (Writing Numbers); within the Language domain, Interview and Interests loaded .82 and .78, respectively. The correlation between the Adaptive and Language factors was .84 after the removal of tasks 1 (Cubes)

TABLE 5

Weighted Least Squares Estimates of Factor Loadings, Measurement Errors and Factor Correlation, and Standard Errors (in parenthesis) for Two-Factor and One-Factor Models for the Gesell School Readiness Screening Test

Task	Two-Factor			One-Factor	
	Loading on Adaptive	Loading on Language	Error	Loading on Developmental Age	Error
Gesell School Readiness Screening Test (Eight Tasks)					
1. Cubes	.67(.03)	---	.56(.08)	.66(.03)	.57(.08)
2. Copy Forms	.85(.02)	---	.28(.07)	.84(.02)	.30(.07)
3. Incomplete Man	.80(.02)	---	.37(.07)	.80(.02)	.36(.07)
4. Writing Name	.86(.02)	---	.26(.07)	.85(.02)	.27(.07)
5. Writing Numbers	.89(.02)	---	.20(.07)	.88(.02)	.22(.07)
6. Interview	---	.78(.03)	.38(.08)	.74(.02)	.45(.07)
7. Animals	---	.73(.03)	.46(.08)	.69(.03)	.52(.07)
8. Interests	---	.86(.03)	.27(.08)	.82(.02)	.33(.08)
Gesell School Readiness Screening Test (Six Tasks)					
2[a]. Copy Forms	.85(.02)	---	.28(.07)	.84(.02)	.29(.07)
3. Incomplete Man	.78(.02)	---	.39(.07)	.78(.02)	.38(.07)
4. Writing Name	.86(.02)	---	.26(.07)	.86(.02)	.26(.07)
5. Writing Numbers	.91(.02)	---	.18(.07)	.90(.02)	.18(.07)
6. Interview	---	.82(.03)	.32(.08)	.74(.02)	.45(.07)
8. Interests	---	.78(.03)	.39(.08)	.70(.02)	.51(.08)

Note. Correlation between Adaptive and Language for the eight task Gesell School Readiness Screening Task was .85 (standard error =.02). Correlation between Adaptive and Language for the six task Gesell School Readiness Screening Task was .84 (standard error =.03).

[a]Original task numbers were used to represent the six tasks in the Gesell School Readiness screening Task.

and 7 (Animals). The 95% confidence interval for this correlation was .78 to .90, indicating a very strong relationship between the constructs, but not a complete overlap. In the one-factor model, shown in the lower half of Table 5, loadings were strong and ranged from .70 (Interests) to .90 (Writing Numbers).

DISCUSSION

The major purpose of this study was to use Rasch analysis and confirmatory factor analysis to investigate the dimensionality of an early childhood test (GSRT), taking into account the theoretical basis of scale construction. Earlier studies of the GSRT using EFA had yielded an acceptable one-factor solution, but had also pointed to the possibility of two factors that corresponded to the Adaptive and Language subdomains of the test. A secondary purpose of this study, therefore, was to evaluate the internal properties of the test, and consider implications of the findings for test score use.

Theoretical issues examined in this paper suggested that the applicability of factor analytic or Rasch psychometric techniques for examining dimensionality should be decided based upon the purposes of scaling and the processes used to operationalize the construct. Thus, the question should be asked, was the intent to scale items or persons, or both, on a continuum? Alternatively, one could ask, was a domain-sampling approach used in developing the instrument, or a Thurstone-type approach, i.e., was there an attempt to order the items by difficulty on the hypothesized scale during test construction?

The GSRT appeared to have been developed using a combination of the above approaches. The tasks of the GSRT are hypothesized to define a global construct, Developmental Age, which is comprised of two subdomains of tasks. The tasks also have a multi-step difficulty structure built into their scoring scheme. Thus, it seemed reasonable to explore the internal properties of the test using both CFA and Rasch analysis.

The Rasch analysis indicated that the data from six of the eight tasks yielded a unidimensional scale as defined by the partial credit model. However, this combination of tasks crossed over the two logical subdomains of the GSRT, Adaptive and Language behaviors. In other words, although the Rasch results pointed to a unidimensional structure of the test, this dimension was not consistent with either of the two theoretical subdomains of the GSRT. When the separate subdomain tasks of the GSRT were independently examined for unidimensionality, misfit continued to surface

for one task. Six tasks appeared to define a single variable that was consistant with the Rasch model.

It was possible to verify using Rasch analysis that response category probability curves (item characteristic curves) were not the same across all tasks of the GSRT. Causes for item and person misfit could also be investigated in depth using Rasch analytic techniques. For instance, the Rasch analysis was found to be sensitive to the frequency with which each scoring category was used in the GSRT tasks, a factor that resulted in some of the observed misfit.

The findings of the CFA using all eight tasks supported the theoretical two-factor structure of the test, in that the data were verified to have acceptable fit to the model with Adaptive and Language task clusters. However, the two factors were found to be highly correlated, which supported the test developers' claim that, together, they comprise the overall construct of Developmental Age. Although the fit of the data to the one-factor model was not as good as that to the two-factor model, the CFA findings were generally consistent with the results of the EFA study cited previously.

Repeating the CFA analysis using tasks that the Rasch analysis defined as unidimensional, reconfirmed the initial CFA findings, even with the smaller subset of six tasks. The Adaptive and Language clusters were again supported and found to be very strongly correlated. The two-factor model also continued to have better fit than the one-factor model.

In terms of implications for test score use, the CFA results supported the use of both the overall DA score, as well as the scores from Adaptive and Language clusters. Although the latter were found to provide some unique information, recommendations for separate educational programming of students based on differential performance on the subtests, should be avoided because of their strong correlation. Use of the GSRT scores for developmental diagnosis or placement would require examinations of predictive validity and classification accuracy, both of which were outside the scope of the present study.

The CFA technique was not helpful in examining the location of tasks by difficulty on a continuum, nor at examining in detail the step difficulty structures within each task. The Rasch analysis provided insights into such details. Rasch models impose rather stringent requirements on the data, resulting in task or person misfit when the data depart from the expectations of the model --" an item either measures the latent trait or it does not, even if it measures something highly

correlated with the latent trait. ...the.. formulation does not allow the latent trait measured by the item to be represented by a composite of two or more latent traits" (Duncan, 1984, p. 386).

In applying Rasch models to determine unidimensionality of test data, several factors should be considered. Choice of the right mathematical model is critical, as was seen in earlier phases of the analysis of the GSRT data when the rating scale model was used. Smith (1996) distinguishes between theoretical and functional dimensionality of tests. The conditions under which the data are collected, which would be categorized under factors affecting functional aspects of test dimensionality, influence the extent to which the data fit the psychometric models used. Nonsensical responses yield poor or inadequate fit, no matter what model is used to detect dimensionality. The more control there is for error in measurement during the data collection process, the better the data will lend themselves for an examination of dimensionality. Ambiguous item wording, unexpected response sets ("Christmas treeing"), untrained or inconsistent scorers / scoring systems, and poor directions, are some factors that may yield misleading results (Duncan, 1984).

The different findings of the CFA and Rasch analysis of the GSRT may be rooted in their different approaches. While CFA is concerned with estimated covariances among sets of items, the Rasch approach focuses on estimates of an item parameter, namely, the item difficulty parameter. For unidimensionality to be manifested, it is required that similarly discriminating items are ordered on a continuum of difficulty. The stringency of the Rasch approach may lead to elimination of tasks or items from a test that may substantively define its content domain.

Rasch models, then, are most useful when the test developer is certain that the process of test construction was deliberately designed to yield an ordering of items that fits the specifications of the Rasch model. Properties of item order or discrimination are not well detected by correlational techniques such as factor analysis. Both EFA and CFA yield useful information on the internal covariance structure of items comprising a test. To avoid misleading results, selection of analytic tools for examining psychometric properties of a test should be done carefully and judiciously.

ACKNOWLEDGEMENTS

The authors thank Benjamin D. Wright for many useful suggestions and comments on an earlier version of this article.

REFERENCES

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

Banerji, M. (1992). Factor structure of the Gesell School Readiness Screening Test. *Journal of Psychoeducational Assessment, 10,* 342-354.

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

Bentler, P. M. (1992). On the fit of models to covariances and methodology to the *Bulletin. Psychological Bulletin, 112,* 400-404.

Browne, M. W., & Mels, G. (1990). *RAMONA PC: User's manual (Ver. 3.0).* Pretoria: University of South Africa.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York, NY: CBS College Publishing.

Duncan, O. D. (1984). Rasch measurement: Further examples and discussion. In C.F. Turner and E. Martin (Eds.), *Surveying subjective phenomena Vol. 2* (pp. 367-403). New York, NY: Russell Sage Foundation.

Hoyle, R. H. (1995). The structural equation modeling approach: Basic concepts and fundamental issues. In R. H. Hoyle (Ed.), *Structural Equation Modeling: Concepts, Issues, and Applications* (pp. 1-15). Thousand Oaks, CA: Sage Publications.

Ilg, F. L., Ames, L. B., Haines, J., & Gillespie, C. (1978). *School readiness.* New York, NY: Harper & Row.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 57,* 409-426.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8 user's reference guide.* Chicago, IL: Scientific Software International, Inc.

Long, J. S. (1983). *Confirmatory factor analysis: A preface to LISREL.* Beverly Hills, CA: Sage Publications.

Marsh, H. W. (1987). The factorial invariance of responses by males and females to a multidimensional self-concept instrument: Substantive and methodological issues. *Multivariate Behavioral Research, 22,* 457-480.

Marsh, H. W., Balla, J. R., & McDonald, R. P (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391-410.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (Third Edition)* (pp.13-103). New York, NY: American Council on Education and Macmillan Publishing Company.

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indices for structural equation modeling. *Structural Equation Modeling, 3(4),* 369-379.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement, 51,* 541-565.

Smith, R. M. (1992). *Assessing unidimensionality for the Rasch rating scale*

model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically-based tests for the number of common factors*. Paper presented at the annual Spring meeting of the Psychometric Society, Iowa City, IA.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,*273-286.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitudes*. Chicago, IL: University of Chicago Press.

Wright, B. D. (1967). *Sample free test calibration and person measurement*. Paper presented at the ETS Invitational Conference on Testing Problems, Princeton, NJ.

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

Wright, B. D., & Linacre, J. M. (1996). *BIGSTEPS: Rasch analysis for all two-facet models*. Chicago, IL: MESA Press.

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

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

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.