Volume 4, Number 1, 2000 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 4, Number 1 2000

REVIEWER ACKNOWLEDGEMENT

The Editor would like to thank the members of the Editorial Board who provided manuscript reviews for the Journal of Outcome Measurement, Volume 4, Number 1.

Editorial

Transition or Change?

Transition or change causes much of our environment to look and act a little different with each advancing hour, day, etc. We as a society keep looking for these differences in antiques, family trees, celestial events, oceanic inhabitants and archeological digs. The question at hand is how the transition or change, which cause these differences, will affect us.

The Journal of Outcome Measurement is not exempt from transition or change. Dr. Richard Smith was instrumental in the birth of the JOM for its parent the Rehabilitation Foundation, Inc. also d.b.a. Rehabilitation Frontiers International. We at RFI thank him for his leadership and foresight. Dr. Smith is no longer with RFI or JOM and has embarked on a new venture with the establishment of the Journal of Applied Measurement. We applaud this effort and welcome the scientific thought and studies that will be provided by Dr. Smith and the authors publishing in this new journal.

But what about the JOM? It will be different. Please read the Concept and Rationale sections of the Contributor Information page. We will transition the solicitation of manuscripts to include more clinical and programmatic outcome studies using the scientific measurement methodologies as published currently and previously in the JOM and elsewhere. Thus the JOM will transition, not suddenly change. I invite measurement AND clinical scientists or any research scientist studying outcome concepts to submit their manuscripts for publication in the JOM. Additionally, your suggestions for the JOM, will be most appreciated during this transition and in the future.

The measurement and study of outcomes in the health and social

sciences is a key science that helps analyze transition or change. Transition is always time laborious where change may be instantaneous. In some circumstances the two seem indistinguishable. That is when outcome measurement becomes invaluable. The science of outcome measurement and studies is critical to our societal need to survive with quality and longevity. We must understand what is happening in order to understand how we can or cannot influence transition or change, in order to achieve a successful end or goal. This may apply to a person recovering from a stroke, to the institution reorganizing to improve financial viability or to our society losing control over behaviors of our youth. In order to provide information to assist in correcting or improving the outcome, the methods must include repeated measurements to determine whether the outcome was the result of transition or sudden change.

Recently, an outcry occurred when a large number of grade school students were informed that they would not enter high school because of poor scores on a standardized test. Many of these students had excelled in their classes and at least one was an honor student. Such examples of single "point in time measurements" to show progress or achievement are numerous in the health and social sciences. Most of the time these "point in time measurements" do not answer the question as to whether the outcome was the result of transition (or lack thereof) factors or sudden change (or lack thereof) factors. Without such information it is difficult to determine how to improve the processes to meet outcome goals.

In my opinion, it is time to expand our vision regarding outcomes using scientifically measurable methods, over time. This, so that we may improve the processes of stroke recovery, improve the operation of the institution and intervene to prevent destructive behaviors from damaging our most important resource—our youth. Report cards, profiling, comparing admission to discharge gains, benchmarking, etc. may not be enough.

This issue of the JOM includes several articles presented at the Sec-

ond International Outcome Measurement Conference held at the University of Chicago May 15 and 16, 1998. These articles bring to you, the readers, some important concepts of measurement.

Please enjoy your reading and please consider submitting, to the JOM, manuscripts on outcome methods and studies that you would like to share with the clinical and scientific community.

Please call: 630-221-1200 x222
Or fax: 630-221 1201
Or E-mail: rfharvey@rfi.org.
JOM Editor,

Richard F. Harvey, M.D.

JOURNAL OF OUTCOME MEASUREMENT®, 4(1), 413-430

Assessment of Unidimensionality of Physical Functioning in Patients Receiving Therapy in Acute, Orthopedic Outpatient Centers

Dennis L. Hart

Focus On Therapeutic Outcomes, Inc., Knoxville, TN

Physical functioning is a common construct of interest for patients receiving rehabilitation. This report describes the assessment of hierarchial structure, unidimensionality and reproducibility of item calibrations along the continuum of physical functioning defined by the PF-10 of the MOS SF-36. Three new questions specific to patients with upper extremity impairments were added, and item calibrations were compared across several groups of patients with different musculoskeletal impairments. Reproducibility of item calibrations over testing times was supported. Item order was dependent on impairment in a clinically logical pattern. Construct validity of the physical functioning scale was supported and improved with the new questions for patients with upper extremity impairments as well as for patients with some lower level extremity impairments.

Requests for reprints should be sent to Dennis L. Hart, Focus on Therapeutic Outcomes, Inc., 551 Yopps Cove Rd., White Stone, VA 22578, e-mail: hart@fotoinc.com

People who seek rehabilitation in acute orthopedic outpatient centers commonly receive services designed to improve physical function and reduce pain complaints associated with musculoskeletal impairments. Recent research has begun the process of describing the patterns of health status outcomes in these patients with common orthopaedic diagnoses (Di Fabio and Boissonnault, 1998), spinal impairments (Jette and Jette, 1996a; Patrick, Deyo, et al., 1995; and Riddle and Stratford, 1998), and knee impairments (Jette and Jette, 1996b). Others (Hart and Dobrzykowski, 2000a) have used health status outcomes to differentiate clinical quality and value between therapists with advanced vs. entry level training who work in these settings.

The construct of physical functioning was a primary concern in these investigations. Physical functioning "refers to performance of a variety of physical activities that are normal for people in good physical health, including activities such as walking, climbing stairs, and carrying objects (Haley, McHorney and Ware, 1994)." The measure of physical functioning described in these articles was the 10-item Physical Functioning Scale (PF-10) from the Medical Outcomes Study (MOS) SF-36, a generic health-related quality of life outcomes instrument (McHorney, et al., 1994; McHorney, et al., 1993; and Ware and Sherbourne, 1992). The PF-10 was designed to assess self-care, mobility and more general physical activities and body movements (Haley, McHorney and Ware, 1994), and therefore appears well suited for patients with musculoskeletal impairments seeking rehabilitation.

Jette and Jette (1996a and 1996b) were the first to describe changes in physical functioning via the PF-10 in patients with lumbar, cervical and knee impairments receiving treatment in acute orthopedic outpatient centers. They reviewed health outcomes in patients in the 1993-1994 Focus On Therapeutic Outcomes (FOTO) database. Using standardized effect sizes (Cohen, 1977), patients reported moderate to large reductions in physical functioning at initial evaluation as compared to the gender- and age-controlled norms from national (USA) surveys (Ware, et al., 1993). Improvement in physical functioning over the course of treatment was described as well with ef-

fect sizes as small to moderate (Cohen, 1977) for patients with spinal impairments and large for patients with knee impairments. Di Fabio and Boissonnault (1998) expanded this work by examining specific diagnoses by ICD-9-CM codes (Puckett, 1995) in patients being treated in similar rehabilitation settings. Comparisons of patients to national norms at intake were generally large effect sizes where the effect sizes over treatment ranged from small to large depending on diagnosis.

However, as described by Tennant and Young (1997), the responsiveness to clinical change of ordinal scores, such as the responses for the PF-10, may be adversely affected by floor and ceiling effects. This raises concern about the appropriateness of effect sizes in these studies, which were calculated from ordinal scales. The use of Rasch transformations (Wright and Stone, 1979; and Wright and Masters, 1982) has been recommended to reduce the effect of any floor or ceiling effects of ordinal measures.

As the assessment of rehabilitation outcomes progresses, the instruments used to quantify outcomes are being scrutinized. Haley, et al., (1994) and Tennant and Young (1997) have extended the investigation of the PF-10 by using Rasch Item measurement techniques (Wright and Stone, 1979). Haley, et al., (1994) used data from the Medical Outcome Study, (Stewart, et al., 1989; Tarlov, et al., 1989) to test unidimensionality and reproducibility of the PF-10 in patients with chronic medical and psychiatric disorders. They described three desirable properties of a multi-item instrument as: hierarchical structure, unidimensionality, and reproducibility of hierarchical structure across different groups of patients and test occasions (Haley, et al., 1994). Hierarchical structure was interpreted as a progression of items from easy to difficult along a clinically appropriate continuum. Unidimensionality refers to the presence of a single dominant construct apparent among different attributes in a multiple item set (Haley, et al., 1994). Reproducibility, in the concept of hierarchical structure, refers to reasonably similar item order and item calibrations along the clinically appropriate mathematical continuum across different groups of patients (sample-free) and test occasions (test-free) (Haley,

et al., 1994).

In the Haley, et al., (1994) analysis, the PF-10 was found to represent a consistent hierarchical order, but that order differed from the hypothesized item order (original test administration of the MOS SF-36 PF-10) (Ware, et al., 1993). The hierarchical order remained relatively consistent across several groups of patients with chronic medical or psychiatric conditions. The item calibrations defined 10 distinct strata (greater than or equal to ±.15 logits as cut-off scores). Although 8 of 10 item-infit statistics exceeded the pre-determined cut-off infit value, the authors stated that the combined results of their analyses supported the PF-10 as a "hierarchical index which is both unidimensional and reproducible" (Haley, et al., 1994).

Table 1

New Physical Functioning Questions

Lifting overhead to a cabinet (Liftover)
Gripping or opening a can (Gripping)
Handling small items such as a pen or coins (Small)

Tennant and Young (1997) examined the PF-10 in patients admitted to a neurological center in England. The largest diagnostic category was multiple sclerosis (16%). It was difficult to compare the degree of severity or chronicity of the patients between the English (Tennant and Young, 1997) and MOS (Haley, et al., 1994) samples. However, using similar Rasch techniques, the fit of the PF-10 data was considered "adequate" in the English sample. The items defined six strata (using the same ±.15 logits as cut-off scores). As with the MOS sample, the hierarchical order of the English sample was different from the original order of administration of the SF-36 PF-10 (Ware, et al., 1993). Comparing the MOS and English samples, the most difficult three items and easiest three items were in similar order, but the order of the center four items was different, which was different than the original order of administration of the SF-36 PF-10 (Ware, et al., 1993). Bathing and dressing showed considerable misfit in both the English and MOS samples.

During preliminary, unpublished (Hart, 1997) investigations of the PF-10 responses in the FOTO data sets, small effect sizes were noted for patients with arm impairments. The relatively small effect sizes were confirmed by Di Fabio and Boissonnault (1998). No data were reported for patients with arm impairments in the English or MOS samples. In an attempt to improve the responsiveness of the PF-10 for patients with arm impairments, three new questions were added to the FOTO physical functioning scale (Table 1). These questions have clinical relevance for patients with arm impairments and use the same three level ordinal responses as the original PF-10 questions. Because of subtle differences in hierarchical order between the MOS and English samples and the need to improve the responsiveness of the PF-10 for patients with arm impairments, the purposes of this investigation were to describe the 1) hierarchical order, 2) unidimensionality, and 3) effect of the new questions on the SF-36 PF-10 in patients with a variety of impairments seeking rehabilitation in acute, orthopedic outpatient centers.

Methods

Patients

63,700 patients (58.5% female 47.9±16.5 yrs, 41.5% male 45.1±15.8 yrs) from 526 clinics from 39 states (USA) treated by 2,738 clinicians during 1997 participated. All clinics were acute, orthopedic outpatient centers, which participated in the Focus On Therapeutic Outcomes, Inc. a national medical rehabilitation database company. The types of outpatient facilities included: 6.6% payer owned, 48.4% hospital, 1.9% physician office, 19.6% private therapist office, 20.9% corporate business, and 2.5% other.

Of the 63,700 patients, the impairments by percent were: 6.1% arm, 15.4% cervical, 28.4% lumbar, 12% hip, 15.7% shoulder, and 15.3% knee. Impairment categories were selected by body part treated. Each patient completed a standardized health status questionnaire at initial evaluation. For a variety of reasons, i.e. patient did not return for discharge, the treating physician discharged the patient prior to discharge, etc., 39,146 patients (61.5%) completed the health status

questionnaire at discharge. The FOTO data collection process has been described (Di Fabio and Boissonnault, 1998; Dobrzykowski and Nance, 1997; Hart and Dobrzykowski, 2000a; Hart and Dobrzykowski, 2000b; Jette and Jette, 1996a; and Jette and Jette, 1996b).

Instrument

The health status questionnaire contained the SF-36 PF-10 questions (McHorney, et al., 19094; McHorney, et al., 1993; and Ware and Sherbourne, 1992) and the three new physical functioning questions (Table 1). Each item is scored 1 to 3 for Yes, Limited A Lot, Yes, Limited A Little, and No, Not Limited At All, respectively. The raw ordinal scores (1-3) were used for the analyses. Previous studies have supported reliability (McHorney, et al., 1994, and Stewart, et al., 1988) and validity (McHorney, et al., 1993, and McHorney, et al., 1992) of the PF-10. Internal consistency for the FOTO PF-10 in this study was similar to previous findings (Hart and Dobrzykowski, 2000a): Cronbach's alpha α=.90 (n=63,700) for intake and α=.91 (n=39,146) for discharge. Internal consistency for the FOTO PF-10 with the three new questions was α=.86 (n=63,7000) for intake and α=.85 (n=39,146) for discharge. The results of this study address validity for the additional three physical functioning questions.

Analysis Techniques and Plan

The data were fit to the Rasch rating scale model by means of the WINSTEPS (Linacre and Wright, 1998) computer program. The Rasch analyses calibrate the patient abilities and the degree of difficulty of the items allowing the relationship between the difficulty of the item and the ability of the patient to be determined. The Rasch model builds a variable continuum based on the responses of the patients in the sample to the items in the instrument, so patients with more ability have a higher probability of giving responses representing fewer limitations as compared to patients with lower ability (Haley, et al., 1994). The model calculates linear measures (in logits) from the raw ordinal scores, which allow inspection of the placement (hierarchical order) and mathematical spacing between the items

along the variable continuum. Inspection allows insight into the theoretical and clinical relevance of the construct of interest, e.g. physical functioning (Haley, et al., 1994).

Thirty different Rasch analyses were performed: intake and discharge (2), all impairment categories grouped and the six different impairment categories (7), with and without the new questions (2), plus intake and discharge for the three new questions for all impairment categories grouped (2) [(2x7x2)+2=30]. Adequacy of fit of each data set to the Rasch model was determined by assessing the infit (mean square information-weighted) and outfit (outlier-sensitive) statistics for the non-extreme scores for patient ability and response difficulty. Fit statistics close to 1.0 mean square were judged to have adequate fit. Hierarchical order (placement along the theoretical/clinical continuum) was assessed through the linear measure (in logits) for each item. Items were described as being in distinct strata if item separation of the standard error of calibration was more than ±.15 logits (Silverstein, et al., 1991). Unidimensionality was assessed by combining the interpretation of item hierarchical order graphically with the interpretation of the individual item infit and outfit statistics, which if between .7 and 1.3 logits indicated adequate fit (Tennant and Young, 1997). Unidimensionality was further interpreted from a clinically relevant perspective, i.e. how the items should be interpreted by impairment specific groups of patients.

Results

Hierarchical Order

Item order from the 28 Rasch analyses (excluding the analyses of just the new questions) could be grouped by hierarchical order into three patterns (Table 2). Pattern #1 is represented by the data set containing patients at discharge from all impairment categories. The order of the items was the same for intake and discharge. The order of the PF-10 questions in the data set for all impairment categories plus the new questions was the same as Table 2 for both intake and discharge. The order of the PF-10 items for patients with lumbar impairments also represented Pattern #1 with the following excep-

Table 2

Fit of SF-36 PF-10 to the Rasch Model at Discharge by Impairment Categories

	Pattern 1 All Categories n=35,366			Pattern 2 Shoulder Impairments n=5,668			Pattern 3 Knee Impairments n=5,919		
Item	Logit	Infit	Outfit	Logit	Infit	Outfit	Logit	Infit	Outfit
Vigorous activities	3.29	0.99	1.81	3.51	0.92	2.21	3.53	1.19	2.43
Moderate activities	1.07	1.05	1.02	1.86	1.02	1.18	0.12	0.94	0.85
Walk over mile	0.75	0.93	0.83	-0.15	1.00	0.80	1.15	0.90	0.84
Bending/kneeling	0.56	1.07	1.06	-0.44	1.04	0.95	2.01	1.15	1.32
Climb several stairs	0.38	0.88	0.79	-0.37	0.94	0.79	1.25	0.93	0.92
Lifting groceries	0.00	1.16	1.11	0.96	1.03	1.03	-1.41	0.99	0.90
Walk several blocks	-0.49	0.80	0.62	-1.10	0.80	0.48	-0.19	0.84	0.69
Climb one flight	-1.37	0.81	0.64	-1.71	0.79	0.48	-0.94	1.00	0.95
Walk one block	-1.91	0.83	0.55	-2.17	0.87	0.41	-2.02	0.84	0.62
Bathing and dressing	-2.29	1.62	2.41	-0.39	1.60	1.62	-3.85	1.43	2.57
Mean		1.01	1.08		1.00	0.99		1.02	1.21
SD		0.23	0.56		0.22	0.53		0.18	0.67

Infit and Outfit measures in mean squares

tion: the order of the items bending/kneeling and walking over a mile was reversed (logit separation =.04 for data set without new questions). The item order was the same for patients with lumbar impairments for intake and discharge. Therefore, the item order for patients with lumbar impairments was similar for all patients regardless of impairment category. Item order was stable over testing times (intake and discharge) and was not affected by the addition of the new questions for Pattern #1.

Pattern #2 is represented by the data set containing patients with shoulder, cervical or arm impairments (Table 2). For each of these impairment categories, the item order for the PF-10 questions was the same for intake and discharge. There was one exception: patients with cervical impairments had the items bathing/dressing and walking one block reversed (logit separation=.1 for the data set with new questions). There were subtle differences in item order between data sets per impairment category for just PF-10 vs. PF-10 plus new questions. Pattern differences of item order for patients with shoulder and arm impairments were the same with the items climbing several flights of stairs, bathing/dressing and bending/kneeling changing order between data sets with or without new questions (logit separation across the three items was .07 for shoulder and .22 for arm impairment groups).

Item order for patients with cervical impairment was different between data sets with and without new questions for items bathing/dressing and walking one block (logit separation=.1 for data set with the new questions). In general, each difference in item order was associated with poor separation (calibration) between items. Overall item calibration implies stability in item order across impairment categories and over different testing times.

Pattern #3 was represented by the item order of patients with lower extremity impairments. Item order between the hip/ankle and knee (Table 2) impairment categories with or without the new questions was generally the same. The groups of items 1) bending/dressing, walking over a mile, and climbing several flights of stairs and 2) climbing one flight and lifting groceries had different item orders (logit separations >.15). The patterns were similar between intake and discharge data sets with or without the new questions. Therefore, although there were some differences in item order, the overall pattern of item order was reasonably similar across impairment categories and was stable between testing times.

Item Calibration

Separation of item standard error of calibrations of .15 logits or more allows determination of the number of distinct strata represented by the questions. The number of strata in each of the 28 data sets varied from 8 to 12. For the 14 data sets with the original PF-10 questions, the number of strata ranged between 8 and 10, and for the 14 data sets with the PF-10 plus the three new questions, the range of strata was 8 to 12.

Unidimensionality

Patients with shoulder impairments and knee impairments (Table 2) offer representative groups for the assessment of unidimensionality. Both groups would be expected to respond to the generic questions of the SF-36 PF-10 differently, since the PF-10 is heavily weighted towards activities requiring lower extremity and overall mobility effort. This clinically relevant difference does not appear to influence the infit statistics in Tables 2, which supports unidimen-

sionality. However, there are subtle differences. There is a more even distribution of item placement for patients with knee impairments as compared to patients with shoulder impairments (item separations). There is some misfitting of the least able patients with knee impairments performing bathing and dressing activities (infit 1.43). However, because the item calibration (pattern) and infit and outfit statistics were good, the PF-10 appears to display unidimensionality. It is clinically logical to expect some degree of misfit for bathing and dressing for these patients as compared to the activities that stress the lower extremities, e.g. walking or climbing stairs. Linacre, et al., (1994) reported that some misfitting of items at the extremes of patient abilities as compared to items representing intermediate patient abilities should be expected.

For patients with shoulder impairments, the activities of climbing several flights of stairs, bathing/dressing, and bending, kneeling, or stooping do not display separate strata (Table 2), although the infit statistics are generally good. The item bathing or dressing shows some misfit for patients with intermediate abilities. From the per-

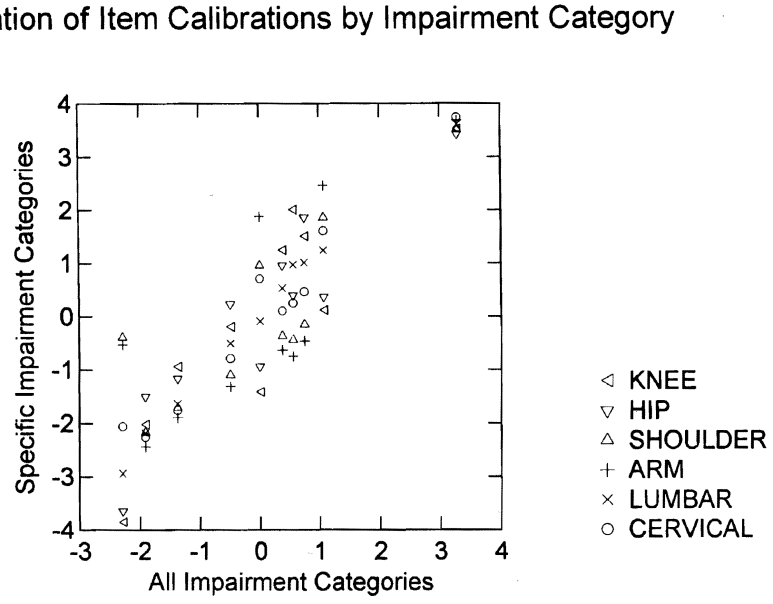
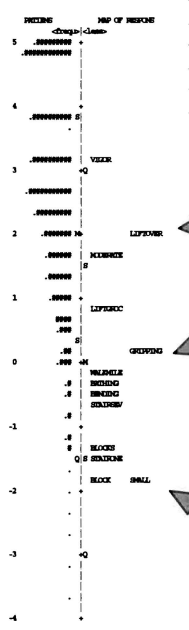


Figure 1. Relation between item calibrations across impairment categories.



Item Calibrations and Fit Statistics For Patients With Shoulder Impairments Including New Questions

Response	Measure	Infit	Outfit
VIGOR	3.19	.84	1.44
LIFTOVER	*1.96*	*1.10*	*1.35*
MODERATE	1.66	.87	.85
LIFTGROC	.81	.88	.79
GRIPPING	*.08*	*1.18*	*1.06*
WALKMILE	-.17	1.13	.92
BENDING	-.27	1.10	1.00
STAIRSEV	-.30	1.06	1.03
BATHING	-.40	1.26	1.17
BLOCKS	-1.36	.88	.48
STAIRONE	-1.55	.83	.48
SMALL	*-1.77*	*1.37*	*1.13*
BLOCK	-1.90	.89	.44

Figure 2. Effect of new physical functioning questions on item calibration of the SF-36 PF-10 for patients with shoulder impairments. Vertical graph (left) from WINSTEPS Table 1.2. Item calibration, measure (in logits) and fit statistics (mean squares) from WINSTEPS Table 13.1. Response is the abbreviated item content.

spective of unidimensionality, these items are "grouped" in the graphic display of patient ability and item difficulty. Unidimensionality of the PF-10 is less well supported for patients with shoulder impairments as compared to patients with knee impairments, but taken together, the knee and shoulder data sets support unidimensionality of the PF-10, with stronger support for the patients with knee impairments.

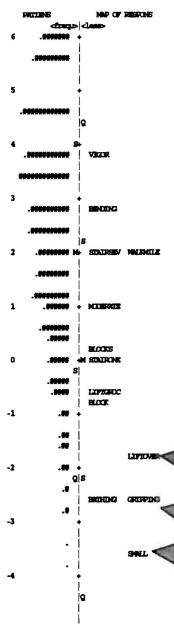
Although the patterns of the graphic displays and the infit statistics may be slightly different between the two examples, the other impairment categories including all impairment categories grouped together support the concept of unidimensionality of the PF-10. Therefore, taken together, the data support unidimensionality (Figure 1).

Effect of New Questions

The purpose of adding the new questions to the PF-10 was to improve the responsiveness to clinical change of the PF-10 for patients with upper extremity impairments. Rasch analyses were not designed to assess responsiveness. However, insight into change in item calibration and unidimensionality can be assessed. The graphic display (Figure 2) for patients with shoulder impairments is a representative example of the influence of the new questions. The infit statistics are adequate, and the display of patient ability and item difficulty demonstrates improvement over the PF-10. The items lifting overhead to a cabinet and gripping or opening a can, improve the distribution of the items, and item order (calibration) is clinically logical. Bathing/dressing, bending or kneeling, and climbing several flights of stairs continue to overlap (logit separation <.15). These items have less clinical relevance for patients with shoulder impairments. Taken together, the results of the Rasch analyses support construct validity.

Of interest, the new questions also improve the calibration of items for patients with knee impairments (Figure 3), in spite of the apparent lack of clinical relevance. There is more misfitting for patients with less ability for gripping or opening a can and handling of small items such as a pen or coins, but the three new questions improve the floor effect of the PF-10 for these patients.

Overall, there was little change in the fit statistics across the impairment categories with the new questions. Item calibration and number of strata identified in each patient group between the PF-10 and the data set with the new questions varied. With all impairment categories grouped together, there was a loss of two strata with the new questions. There was some improvement for patients with arm impairments and shoulder impairments with the ability to differentiate more strata with the new questions, and more strata were identified for patients with lumbar or knee impairments. The pattern for lumbar impairment was similar to the pattern for knee impairment with the new questions improving the floor effect of the instrument. In patients with knee impairments (Figure 3), the addition of the new questions produced a pair of items, walking more than a mile and climbing



Item Calibrations and Fit Statistics For Patients With Knee Impairments Including New Questions

Response	Measure	Infit	Outfit
VIGOR	3.83	1.01	1.63
BENDING	2.72	1.01	1.26
WALKMILE	2.07	.92	1.04
STAIRSEV	2.05	.88	1.13
MODERATE	.92	.86	.80
BLOCKS	.21	.82	.67
STAIRONE	.05	1.00	.88
LIFTGROC	-.54	.98	.80
BLOCK	-.84	.83	.61
LIFTOVER	*-1.71*	*1.19*	*1.51*
BATHING	-2.54	.99	.99
GRIPPING	*-2.70*	*1.60*	*2.16*
SMALL	*-3.51*	*1.77*	*4.10*

Figure 3. Effect of new physical functioning questions on item calibration of the SF-36 PF-10 for patients with knee impairments. Vertical graph (left) from WINSTEPS Table 1.2. Item calibration, measure (in logits) and fit statistics (mean squares) from WINSTEPS Table 13.1. Response is the abbreviated item content.

several flights of stairs, that did not separate well (logit <.15). This should initiate discussion of the possibility of deleting one of the now redundant items, which is encouraged.

Discussion

The data sets fit the Rasch model well, which allows us to use these analyses to discuss the purposes of the investigation. The first purpose was to describe the hierarchical order of the PF-10 across impairment categories in patients receiving therapy in acute orthopedic outpatient centers. Our item order was different from the administrative order of the original MOS SF-36 PF-10 (Ware, et al., 1993). The order was different from the order published by Haley, et al., (1994) for patients with chronic medical and psychiatric conditions and the

patients published by Tennant and Young (1997) for neurologic conditions. However, the patterns of item order from the three data sets had some general similarities. Consider our data set with all impairment categories grouped together. Vigorous activities and bathing/dressing commonly represented the most difficult and easiest activities, respectively. Climbing one flight of stairs and walking one block were slightly more difficult than bathing/dressing. Haley, et al., (1994) and Tennant and Young (1997) reported the second most difficult item to be climbing several flights of stairs. Moderate activities were the second most difficult task for patients seeking rehabilitation. Walking over a mile was the third and bending/kneeling was the fourth most difficult tasks in all three data sets. Therefore, the three data sets lend support to the stability of hierarchical order across diverse patient samples, which lends support for the PF-10 being sample-free.

If patients seeking rehabilitation are separated into their respective impairment categories, three item patterns were evident that appear to have clinical relevance. It is clinically logical that patients with cervical, shoulder or arm impairments would respond differently to the PF-10 questions compared to patients with knee or hip impairments. For example, many people with upper extremity impairments should have little functional difficulty with walking, climbing stairs, or bending/kneeling. These same activities should be of functional concern to many patients with lower extremity impairments. Patients with lumbar impairments have different functional concerns and limitations when compared to patients with upper or lower extremity impairments. The patients with lumbar impairment were grouped, by pattern of item order, to the group of all patients regardless of impairment category.

When you examine all patients seeking rehabilitation together, regardless of impairment category, there are surprising similarities in hierarchical order across the diverse patient samples: chronic medical or psychiatric conditions (Haley, et al., 1994), patients with neurological impairments (Tennant and Young, 1997), and patients seeking therapy in acute orthopedic outpatient centers. Therefore, the data lend support that the PF-10 is "sample free". Because the patterns of item order were consistent per impairment category over intake and discharge, the data confirm that the PF-10 is test-free. Taken together, the data support the

reproducibility of the PF-10 in these patient samples.

Because there were three clinically relevant patterns of hierarchical order when patients were separated into their respective impairment groups, our data support the construct validity of the PF-10 as a measure of physical function. For an instrument to demonstrate validity, it must measure what the designers intended it to measure. The PF-10 was intended to measure physical functioning across a broad spectrum of physical functioning constructs (McHorney, et al., 1994; McHorney, et al., 1993; and Ware and Sherbourne, 1992). For the PF-10 to be valid, less difficult items should correspond to activities that are observed to be clinically easier to perform. The same should be true for more difficult activities (Linacre, et al., 1994). This hypothesized order of items per patient pattern is present and is clinically relevant. Therefore, "statistical validity" (Linacre, et al., 1994) is evident in our fit statistics (Table 2), which supports construct validity of the PF-10 for patients in rehabilitation.

The second purpose was to assess unidimensionality of the PF-10 in patients seeking rehabilitation. Although unidimensionality is probably never completely confirmed (Haley, et al., 1994), our data support the presence of unidimensionality. The graphic patterns of distribution of the items and infit statistics for patients with knee, lumbar and cervical impairments and all patients grouped regardless of impairment category demonstrate stable distributions. Infit statistics were generally in the pre-determined range of .7 to 1.3 logits. There were some items that misfitted, but the misfit was clinically logical.

The third purpose was to determine the effect of combining three new questions with the MOS SF-36 PF-10. The results confirm an improvement in the item calibration, strata identification and unidimensionality of the new physical functioning scale for patients with arm, shoulder, lumbar and knee impairments. Since the questions were written to address the functional concerns of the PF-10 in patients with upper extremity impairments, we were surprised to see the improvement in the characteristics of the instrument for patients with lower extremity impairments. The difference between the item calibration in upper vs.

lower extremity patient groups is evident in the location of the items in the theoretical continuum. The new questions assisted in lowering the floor of the instrument for patients with lower extremity impairments. The new questions improved the middle of the ability continuum for patients with upper extremity impairments. This difference in item calibration between impairment groups may be related to the level of difficulty patients perceived when answering the questions. Patients with upper extremity impairments perceived the new items to have intermediate difficulty as compared to patients with lower extremity impairments who perceived the questions as easy. The questions confirm the tendency of the physical functioning scale to be sample or impairment dependent in patients receiving rehabilitation in acute orthopedic outpatient centers.

The next step is to reduce the number of questions necessary to assess physical functioning in the samples of patients of interest, and explore the need for additional questions to reduce the effects of floor and ceiling effects. The Rasch techniques will continue to be of value in this process.

References

Cohen, J. (1977). *Statistical Power Analysis for the Behavior Sciences*. New York, NY: Academic Press Inc., 1977.

Di Fabio, R. P., and Boissonnault, W. (1998) Physical therapy and health-related outcomes for patients with common orthopaedic diagnoses. *Journal of Orthopaedic and Sports Physical Therapy. 27*(3), 219-231.

Dobrzykowski, E. A., and Nance, T. (1997). The Focus On Therapeutic Outcomes (FOTO) outpatient orthopedic rehabilitation database: results of 1994-1996. *Journal of Rehabilitation Outcomes Measurement, 1*, 56-60.

Haley, S. M., McHorney, C. A., and Ware, J. E. (1994). Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): I. Unidimensionality and reproducibility of the Rasch Item Scale. *Journal of Clinical Epidemiology*, *47*(6), 671-684.

Hart, D. L., and Dobrzykowski, E. (2000a). Impact of exercise history on health status outcomes in patients with musculoskeletal impairments.

In Schunk C (ed). Functional Outcomes. Orthopaedic Physical Therapy Clinics of North America. Philadelphia, PA: W.B. Saunders Co.; 9(1):1-16.

Hart, D. L., and Dobrzykowski, E. (2000b). Influence of orthopedic clinical specialist certification on clinical outcomes. *Journal of Orthopaedic and Sports Physical Therapy*, 30, 183-193.

Jette, D. U., and Jette, A. M. (1996a). Physical therapy and health outcomes in patients with spinal impairments. *Physical Therapy*, *76*(9), 930-945.

Jette, D. U., and Jette, A. M. (1996b). Physical therapy and health outcomes in patients with knee impairments. *Physical Therapy*, *76*(11), 1178-1187.

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

Linacre, J. M., & Wright, B.D. (1998). *A User's Guide to WINSTEPS: Rasch-Model Computer Program*. Chicago, IL:MESA Press.

McHorney, C. A., Ware, J. E., Lu, J. F. R., and Sherbourne, C. D. (1994). The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care, 32*, 40-66.

McHorney, C. A., Ware, J. E., and Raczek, A. E. (1993). The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care, 31*, 247-263.

McHorney, C. A., Ware, J. E., Rogers, W., Raczek, A. E., and Lu, J. F. R. (1992). The validity and relative precision of the MOS short- and long-form health status scales and Dartmouth COOP charts: results from the Medical Outcomes Study. *Medical Care, 30*(Suppl.), MS253-265.

Patrick, D. L., Deyo, R. A., Atlas, S. J., Singer, D. E., Chapin, A., and Keller, R. B. (1995). Assessing health-related quality of life in patients with sciatica. *Spine*, *20*, 1899-1909.

Puckett, C. D. (1995). The Educational Annotation of ICD-9-CM. 4th Edition. Channel Publishing: Reno, NV.

Riddle, D. L., and Stratford, P. W. (1998). Use of generic versus region-specific functional status measures on patients with cervical spine disorders. *Physical Therapy, 78*, 951-963.

Silverstein, B., Kilgore, K. M., Fisher, W. P., Harley, J. P., and Harvey, R. F. (1991). Applying psychometric criteria to functional assessment in medical rehabilitation: I. Exploring unidimensionality. *Archives of Physical Medicine and Rehabilitation, 72*, 631-637.

Stewart, A. L., Greenfield, S., Hays, R. D., Wells, K., Rogers, W. H., Berry, S. D., McGlynn, E. A., and Ware, J. E. (1989). Functional status and well-being of patients with chronic conditions: Results from the Medical Outcomes Study. *Journal of the American Medical Association, 262,* 907-913.

Stewart, A. L., Hays, R. D., and Ware, J. E. (1988). The MOS Short-Form General Health Survey: reliability and validity in a patient population. *Medical Care, 26*(7), 724-732.

Tarlov, A. R., Ware, J. E., Greenfield, S., Nelson, E. C., Perrin, E., and Zubkoff, M. (1989). The Medical Outcomes Study: An application of methods for monitoring the results of medical care. *Journal of the American Medical Association, 262*, 925-930.

Tennant, A., and Young, C. (1997). Coma to community: Continuity in measurement. In Smith, R. M. (editor). *Outcome Measurement: State of the Art Reviews. Physical Medicine and Rehabilitation, 11*(2), 375-384.

Ware, J. E., and Sherbourne, C. D. (1992). The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Medical Care, 30*, 473-483.

Ware, J. E., Snow, K. K., Kosinski, M., and Gandek, B. (1993). *SF-36 Health Survey Manual and Interpretation Guide*. Boston, MA: The Health Institute, New England Medical Center.

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

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

JOURNAL OF OUTCOME MEASUREMENT®, 4(1), 431-452

Identifying Shortcomings in the Measurement of Service Quality

Gerard Fogarty, Ph.D., MAPS
The University of Southern Queensland

R. Catts

C. Forlin

SERVPEFR, the performance component of the Service Quality Scale (SERVQUAL), has been shown to measure five underlying dimensions corresponding to Tangibles, Reliability, Responsiveness, Assurance, and Empathy (Parasuraman, Zeithaml, & Berry, 1988). This paper describes three separate studies employing SERVPERF in an Australian context. In the first of these studies (N=113), a shortened 15-item version of the SERVPERF scale (SERVPERF-R) was found to be suitable for use in an Australian small business setting. A five-factor structure was identifiable but the factors were highly correlated, suggesting that they were not clearly distinct. The tendency for marked negative skewness observed by other researchers was also noted here. A follow-up study involving three other small businesses (N=212) used Rasch analysis to test assumptions about the spread of items on the underlying continuum. These analyses indicated that there is an even, though narrow, spread of items across the continuum. The Rasch analysis suggested that the items in both SERVPERF and SERVPERF-R are too easy to rate highly and that more "difficult" items need to be added to the scale. The third study (N=122) was conducted using a version of SERVPERF-R that included seven new items intended to extend the range of the scale. The new items, however, did not achieve this desirable outcome. The implications for service quality assessment are discussed.

Requests for reprints should be sent to Gerard Fogarty, Ph.D., MAPS, The University of Southern Queensland, Toowoomba Queensland , 4350, AUSTRALIA; email: fogarty@usq.edu.au

Measuring Service Quality: Overcoming Ceiling Effects in SERVPERF

The present series of studies reports on the psychometric analysis of one of the main instruments used to measure the construct of service quality, the SERVQUAL scale developed by Parasuraman, Zeithaml, and Berry (1988), and an attempt to extend the range of the scale following the application of Rasch analysis. The psychometric properties of SERVQUAL have been the subject of considerable research in recent times. The scale is said to tap five different underlying dimensions of customer service labelled Tangibles, Reliability, Responsiveness, Assurance, and Empathy that are applicable across a broad range of services. A succession of researchers, however, have failed to replicate the five-factor structure with the number of factors varying from one to five (Finn & Lamb, 1991; Cronin & Taylor, 1992; Gagliono & Hathcote, 1994). Carman (1990) anticipated some of these difficulties when he showed that whilst there are recognizable core dimensions tapped by SERVQUAL, the dimensionality of the scale can change somewhat depending on the type of business being surveyed. A tire business, for example, may be more concerned with courtesy than with security. A 24-hour takeaway business, on the other hand, is likely to place a higher value on security. Carman went on to show that when certain dimensions of quality are perceived as being important by customers, these invariably consist of a number of subdimensions. He suggested that in some industries it is important to add items to cover these subdimensions so that they emerge as fully defined factors in their own right. Thus, his explanation for the differing factor analytic patterns found in the literature was that they can be explained in terms of the context itself with the five major factors retaining their identify or partitioning into others factors depending on the importance of that dimension in the particular context being studied.

Parasuraman, Berry, & Zeithaml (1991) put forward a different explanation when they drew attention to the high scores obtained by most respondents on all subscales of SERVQUAL. They argued that the high scores inflated correlations among all items and made it

difficult to separate the underlying dimensions. Thus, an individual would tend to to use a certain response category (e.g., 7 indicating perceptions of very high service quality) and apply it to most items. Another individual would use a different category (e.g., 5) and apply it to most items. The range of values used by a particular individual was very small, the range across individuals was also small but sufficiently large to permit high inter-item correlations. In a subsequent study designed to explore variations in question format that would overcome the ceiling effect, the same researchers stated that there was a need for further research into the causes of the "upward bias" and ways of correcting the bias (Parasuraman, Berry, & Zeithaml, 1993). Without going into the merits or demerits of Parasuraman et al's (1991) explanation, suffice it to say that the present research programme took up the challenge of examining the problem of upward bias. It did so by employing Rasch analysis to calibrate the items comprising the service quality scale.

Given the uncertain factorial composition of SERVQUAL and what appears to be a strong contextual influence (Carman, 1990) it was considered important to the overall aim of the research to commence by establishing the psychometric properties of the scale in an Australian small-business setting. A preliminary study was conducted for this purpose. The main aim of the preliminary study was to ensure that the subscales of SERVQUAL had high internal consistency and construct validity in an Australian context. A combination of traditional item analysis and exploratory factor analysis (EFA) was used for this purpose.

Method

Participants

A total of 113 customers (21 males) of a small retail business within a provincial city in South East Queensland were surveyed. The business employed both full-time and part-time staff, was owner-managed, and had operated in this area for many years. The majority of staff were long-term employees and some had family links with

the companies. In recent years, the firm had faced increased competition from major retail companies, especially through the development of regional shopping complexes. The firm operated in a strip retail location.

Materials

SERVQUAL consists of 22 pairs of items: one member of each pair assessing the customer's expectations, the other assessing perceptions of service quality. Service quality can be determined by calculating the difference between expectations and perceptions for each item. This aspect of the administration o f SERVQUAL has been criticized on the grounds that the calculation of difference scores results in poor reliability, especially if the expectations scale is truncated by ceiling effects, as is usually the case (Brown, Churchill, & Peter, 1993). The problem of working with difference scores can be overcome by administering the perception questions alone. Cronin and Taylor (1992) found that the perceptions component alone (SERVPERF) out-performed the whole SERVQUAL scale in terms of reliabilities. For this reason, it was decided to use the one-stage (SERVPERF) form of the survey in the present research. Prior to administration, minor wording changes were made to SERVPERF to convert negatively worded items to positive items consistent with the recommendations of earlier researchers (Babakus & Boller, 1991; Parasuraman et al., 1991). Minor changes were also made to some questions to ensure that the text was familiar to Australian consumers.

The version of SERVPEFR used in our first study contained 22 items, most of which were taken unchanged from the perceptions half of SERVQUAL. The first four items in the scale measure Tangibles (e.g., "It has modern looking equipment"), items 5-9 measure Reliability (e.g., "When they promise to do something by a certain time they do it"), items 10-13 Responsiveness (e.g., "They give you prompt service"), items 14-17 Assurance (e.g., "You can trust them"), and items 18-22 Empathy (e.g., "They understand your specific needs").

Procedure

SERVPERF was administered by interview at the point-of-sale by two researchers who were trained in the use of the instrument. Customers were approached only after they had made a purchase. This was to ensure that they were actual customers, not merely "browsers", and therefore familiar with the service they were being asked to evaluate. Customers were asked to give their perceptions of service quality for the 22-items on a 7-point Likert scale using categories ranging from "strongly disagree" (1) to "strongly agree" (7).

Results

The first stage of data analysis involved the use of EFA routines from the SPSS package to check the dimensionality of the full 22-item SERVPERF scale. The exploratory analysis was not intended as a strict test of Parasuraman et al's (1988) model, rather it was used as preliminary check of dimensionality in relatively unconstrained conditions and also as a guide for later item analysis. In accordance with the procedures followed by Parasuraman et al. (1991) in their validation study, the principal axis factoring technique was used with the solution constrained to five factors subjected to oblique (oblimin) rotation. The factor pattern and factor intercorrelation matrix is shown in Table 1.

The five-factor solution accounted for 70% of the variance and coincides with the solution that would have been offered under root one criterion. The first factor was defined by items mostly from the Assurance and Empathy scales, with item 7 (Reliability) and item 10 (Responsiveness) also loading here. The second factor was defined by items 1-4 and is clearly the Tangibles factor. The third factor picked up variance from items 11-13 (Responsiveness) and also from items 14-15 (Assurance) plus item 18 and 22 (Empathy). The fourth factor was defined mostly by items 5-8 (Reliability). The last factor picked up variance from the Reliability and Empathy scales. It is apparent that, apart from Tangibles and Reliability, the five factors identified by Parasuraman et al. (1988, 1991) have not emerged clearly here.

Reliability analyses indicated that all five scales had at least reasonable internal consistency reliability (Cronbach's alpha) with estimates ranging from .69 for the Reliability scale to .86 for the Empathy scale. In all cases, however, it was evident that reliability could be improved if some items were deleted. In a second series of reliability analyses, three criteria were applied to help decide whether an item should be deleted. The first criterion was the impact of the item on the reliability of the scale: items that lowered the alpha esti mates for each scale were considered for removal. The second criterion was the location of the item in the exploratory analysis shown in Table 1: items that did not line up with other items in its scale were considered for deletion. The third criterion was the suitability of an item for use in an Australian context, as judged by feedback from the interviewers who administered the scale.

All these criteria were applied jointly to form a reduced 15-item scale called SERVPERF-R (See Appendix A). Descriptive statistics based on dataset 1 for the reduced scale are shown in Table 2. It can be seen from Table 2 that the internal consistency estimates for the 15-item SERVPERF-R subscales were generally satisfactory. It can also be seen that means for the subscales were high, indicating that customers were happy with all aspects of service. In fact, with the exception of the Reliability subscale, all variables were negatively skewed ($p<.001$). The same feature was evident in the 22-item SERVPERF, so skewness was not introduced as a consequence of reducing the scale from 22 to 15 items.

Summarizing the changes that were made to SERVPERF to improve its construct validity in an Australian context: a) some items were deleted using the criteria described earlier, resulting in a 15-item scale; b) items that were originally reverse-scored were rewritten so that all items had a positive orientation; and c) the wording of some items was changed to make them more understandable in the Australian situation.

Study 2

In Study 1, it was observed that there was a tendency for cus-

tomers to give high ratings across the different dimensions of SERVPERF. The question of item separation has not been adequately addressed in the literature, yet it is an issue that is fundamental to the measurement process itself. When Parasuraman et al. (1991) suggested that items of SERVQUAL were attracting similar rates of endorsement they implied that, for the populations sampled, the items are located in similar positions along the service quality continuum. One of the major aims of the second study was to test this assumption by using Rasch analysis (Rasch, 1960) to calibrate the items. Well-constructed scales employ items that cover the whole dimension in regularly spaced intervals (Wright & Stone, 1979). The main aim of the current analysis was to see whether the items of SERVPERF-R were sufficiently spaced to allow separation of persons on the service quality dimension.

Method

Participants

Customers of three additional retail business (N1=63, N2=75, N3=74) in the same regional city were surveyed, thus forming an additional dataset of 212 cases (76 males) that was used for item calibrations.

Materials

SERVPERF-R, the shortened form (15 items) of the original SERVPERF, was the only instrument used in the survey (Appendix A).

Procedure

As for study 1.

Results

Checking Assumptions for Rasch Analysis

One of the main assumptions of Rasch analysis is that the items form a unidimensinal scale. Although Study 1 showed that SERVPERF-R has a recognizable subscale structure, there was also evidence that it can be treated as unidimensional. The factor intercorrelations for both the long and short forms of the SERVPERF scale (see Table 1) suggested that there is a common service quality dimension underlying all the items in this scale. Other aspects of the data analysis also suggested that this was the case. Firstly, in exploratory factor analyses, the first eigen value accounted for 45% of the variance, more than five times that accounted for by the next eigenvalue. Ratios of this magnitude are generally considered to support assumptions of unidimensionality (Hambleton, Swaminathan, & Rogers, 1991). Secondly, reliability analyses with a single 15-item scale indicated that it had an internal consistency estimate of .92. Thirdly, LISREL analyses (unreported) revealed that a model with all items serving as indicator variables for a single underlying dimension produced indices of fit that were not significantly different than those associated with a five-factor model.

Rasch analysis also assumes that items are equally discriminating. The reliability analyses conducted in the first part of this study resulted in a set of 15 items that all had reasonable item-total correlations (between .29 and .76). Item total correlations can be taken as a crude form of item discrimination index, so this assumption was also deemed to have been met.

Item Location Estimates

The ASCORE program (Andrich, Sheridan, & Lyne, 1991) contains an implementation of the multiplicative binomial that is capable of handling data generated by the response format used in SERVPERF-R. The location estimates for the 15 items in the shortened form of SERVPERF-R are depicted graphically in Figure 1. The items have been grouped into factors to make it easier to see the areas of the service quality continuum tapped by the five factors.

Two things are immediately apparent from this graph. The first is that all items have location estimates that fall between -0.6 and

+0.6, so the range is narrow. The second point to note is that the subscales tend to cover different parts of the continuum. Thus, Empathy and Assurance cover the lower end, Responsiveness the middle, and Reliability and Tangibles the upper end. In practical terms, this means that Empathy and Assurance are endorsed more readily than the other dimensions. They are like the easy items in an ability test. Responsiveness, on the other hand, covers a broad span and is generally a more difficult quality to rate highly. The second item in this scale (Item 8: "They are always willing to help you") is endorsed quite readily but the third item (Item 13: "They are never too busy to respond to your requests") is more difficult. Tangibles and Reliability were certainly the most difficult of the factors to rate highly: endorsement of these items implies endorsement of all other items as well.

Person estimates can also be obtained with a Rasch analysis. The person estimates are independent of the item estimates and provide information about how suitable the scale is for a particular sample. The average person estimate in the present study was 1.24, indicating that most of the customers found it easy to endorse the items in SERVPERF-R emphasizing, once again, the need to add more difficult items to the scale. In fact, because the highest item location was only .47, one could say that the average respondent was located beyond the reach of scale items, a point that is illustrated by the large number of 7's in the raw data. The relationship between the span of scale items and the affectivity of the persons being measured is shown more clearly in the item-person map shown in Table 3. Ideally, the spread of items should be even and should match the span obtained for persons. We can see that this is not the case here. The spread of items is even but does not extend far enough upward: no test items exist to differentiate at the upper end of the service continuum scale. A third study was conducted to determine whether SERVPERF-R could be extended to cover this region.

Study 3

Study 2 confirmed that SERVPERF-R is characterised by lack of differentiation among items, with almost all items attracting high

rates of endorsement. Parasuraman et al. (1991) suggested three different approaches to this problem; a) asking customers to indicate on the same scale where they would place a hypothtical high-quality service company and where they would rate the client company; b) asking customers to treat the high end of the rating scale as representing the level of service they would expect from an ideal company and then rating the client company against this; and c) anchoring the scale's midpoint as the service level expected of a high-quality company and then rating the client company against this standard. All three of these options involve altering the format of SERVQUAL/SERVPERF to provide a framework for the respondent. Because such a approach is conceptually more difficult for the respondent, we decided to work with the same item format but attempt to increase the range of the scale by adding more difficult items. Rasch analysis was again used to calibrate the items.

Participants

A total of 122 customers (27 male) were surveyed from the same firm used for Study 1. The second data collection took place almost 12 months after the first, so it is unlikely that any of the customers were the same.

Materials

The three researchers developed an additional set of 15 items that were considered to be more difficult for customers to rate highly. Unlike the first version of SERVPERF-R, in which negatively worded items were avoided, the new set of items included some items of this type. The 15 new items were then shown to an expert panel which was required to rate the items on the basis of ease of administration (i.e., intelligibility) and also on relative difficulty. As a result of this review process, seven items were selected for inclusion in SERVPERF-R. Three of these items (17, 21, 22) were negatively worded. The additional items, and the factors for which they were judged to be markers, are shown at the bottom of Appendix A.

Procedure

The same procedure was followed as for the previous two studies. That is, customers who had made a purchase were approached by the interviewer and asked to complete this 22-item SERVPERF-R.

Results

To conserve space, only the data relating to the Rasch analyses are reported here. These analyses showed that two of the three negatively worded items did extend the range. Item 17, for example, had a logit value of 1.23. Exploratory factor analysis, however, revealed that the three negatively worded items formed what appear to be a separate "Method" factor rather than defining the traits for which they were selected as markers. The remaining four new items exhibited favorable psychometric properties but had difficulty indices that fell within the range established by the 15-item version of SERVPERF-R. The attempt to extend the range of the scale by adding new items was therefore unsuccessful.

Discussion

The aim of the first study was to explore the psychometric properties of the 22-item version of SERVPERF in an Australian small business setting. This led to the formation of a 15-item version named SERVPERF-R. There is no doubt that the shortened form has acceptable psychometric properties and measures the same underlying traits as the longer version. In fact, LISREL analysis (unreported) suggested that it provided a better fit than the 22-item version (SERVPERF). In order to achieve a good fit for even the shortened form, however, items had to be permitted to load on other factors, with the result that the factors were highly correlated. Parasuraman et al. (1991) made this same observation when noting that SERVPERF has a "diffused" factor pattern and high factor intercorrelations. They argued that the overlap among the dimen-

sions is a function of a tendency for respondents to rate a particular company highly on all dimensions (Parasuraman et al., 1991, p.443). That is certainly what happened in Study 1 with the means for all SERVPERF and SERVPERF-R subscales towards the upper limit.

It would be difficult, however, to establish this as the source of dimensionality problems using conventional item and test validation techniques. Conventional scales do not conform strictly with linearity assumptions and it is not easy to tell just what raw scores mean. Study 2 introduced Rasch analysis to further explore the problem of high rates of endorsement. One very important aspect of a Rasch analysis is that it locates items on a linear continuum that has a zero point and equal units of measurement (logits) extending in either direction from this point. The location estimates are sample free and give an indication of the "ease of endorsement" of an item.

It can be seen from the location estimates shown in Figure 1 that the subscales of SERVPERF-M, if arranged in the order shown in this figure, form a progression. Empathy and Assurance were the easiest to endorse: the businesses in this study found it rather easy to achieve standards in this area. Responsiveness and Reliability were somewhat harder to achieve and Tangibles was the hardest area in which to be rated a success. These outcomes may seem counterintuitive to readers who expect Tanglibles to be a rather easy area in which to achieve service quality. After all, it only involves the outlay of money. In fact, compared to major chains, capital was a challenge for all businesses that took part in this study. They were mostly local, family-owned retail stores facing increasing competition from multinational companies moving into the region. The retail stores look "old-fashioned" when compared with their modern, multinational competitors. Therefore, the difficult aspects of service quality for the local firms were physical facilities (item 6) and modern looking equipment (item 1). What they can supply more readily is product knowledge (item 14) and personal attention (item 10).

Although interesting, because it illustrates another way in which context can influence the measurement of service quality, the observation of an apparent ordering of the subscales on an unidimensional

difficulty continuum was not a central concern of the study. Of more importance was the spread and range of the location estimates. Dealing with the spread first, if one drops a perpendicular line from each item to the baseline of Figure 1, it can be seen that there are no gaps in the scale. SERVPERF-R actually covers the middle and lower sections of the underlying continuum rather well. Despite the even spread of items, however, it would certainly be desirable to extend the range of the scale somewhat with a view to achieving better discrimination among the supposed five latent traits. It will always be difficult to determine the factor structure of SERVPERF or SERVPERF-R while raw scores are so clustered. There is a need for items that respondents will find more difficult to rate highly. Table 3 shows this quite clearly. According to this item-person map, there were a lot of people in this sample whose perceptions of service quality could not be tapped by SERVPERF-R. They registered 7's for almost every item in the scale. It should be noted that none of the items discarded for the shortened version of SERVPERF had a logit score above 0.47. Although not reported, a Rasch analysis was conducted on the full 22-item scale used in the first sample. Six out of the seven discarded items had location values very close to zero. The remaining items did not extend the scale beyond the bounds covered by the shortened version.

The additional items included in SERVPERF-R in the third study were intended to achieve this extension. The items chosen for inclusion were judged by the research team to be more difficult for customers to rate highly. To overcome a possible tendency to respond in a positive fashion to all questions, three of the items were negatively worded. Parasuraman et al. (1991) tried this approach but advised against using negatively worded items because a) they had higher variance than other items, b) they decreased the reliability of subscales, and c) people found these items harder to understand (p. 422). On the basis of the Rasch analysis of data from our third study, it was evident that the use of negatively worded items had the effect of extending the range of the scale. On these grounds, they might seem to be desirable additions to the range of the scale. However,

the problems noted by Parasuraman et al. (1991) in relation to negatively worded items were also observed here. In particular, they lowered the reliability estimates for their subscales. The reason for the lower reliabilities was quite apparent in Study 3: negatively worded items no longer defined the traits tapped by the rest of the items in the subscales but instead combined to define what can best be described as a separate "method" factor. There are many reasons for items combining to define a factor. Rather than speculate on these reasons here, we simply agree with previous researchers (Parasuraman et al., 1991; Babakus & Boller, 1991) that negatively worded items are likely to create some confusion in the minds of customers. Taking into consideration the comments made by Marsh (1996) and Spector (1997) on artifactual factors that can be introduced by negatively worded items, we are inclined to conclude that these items had such an effect in the present instance. With the negatively worded items taken out of the extended SERVPERF-R, it is apparent that the attempt to create more difficult items in this fashion was unsuccessful.

The implications of these findings require careful consideration. On the surface, SERVPERF-R appears to be an adequate instrument: it contains five subscales, each of which has satisfactory internal consistency reliability estimates and each of which can be identified in confirmatory factor analysis. The real problem with SERVPERF-R, and its parent scale SERQUAL, is not that its structure can vary according to the context or that the subscales have poor reliability, but that customers tend to use the highest rating categories (5,6, and 7), with consequent ceiling efforts. The use of Rasch analysis and item-person maps in the present study illustrated this point in a graphic manner. The solution to the problem, however, does not lie in simply trying to develop more difficult items. With the exception of the problematic negatively worded items, we were not able to develop such items in Study 3.

It may be that the solution lies in a combination of the approaches recommended by Parasuraman et al. (1991) and the use of an item calibration technique such as Rasch analysis. Parasuraman et al. recommended that respondents be asked to consider an ideal company

and to see it as providing the midpoint of a scale and then to place the client company somewhere on the same scale. Parasuraman, Zeithaml, and Berry (1994) went on to explore alternative question formats but were not able to solve the problem of "upward bias" by this technique alone. They recommended that "Research aimed at understanding the cause of this phenomenon and estimating the extent of upward bias it produces would be helpful in reducing the bias, or at least correcting for it in interpreting direct-measure ratings" (p.221). Their attempt to introduce a scaling factor at the interview stage is interesting because Rasch analysis achieves a similar rescaling by setting the mean of the item difficulty estimates to zero. We can see the effect of this Figure 1. It is possible that if the question format itself were also changed in the manner indicated by Parasuraman and his colleagues more variation would be observed in both the raw score and logit estimates. That is, in the first instance, introduce some form of anchoring in the scale itself so that customers have a benchmark against which to compare the client company. Secondly, use Rasch analysis to calibrate the items so that they form a true linear scale and then derive person scores from these item estimates. Scores derived in this manner will prove useful for discriminating among companies and for measuring change in performance within a company as a consequence of training or some other intervention. Until some progress is made in this area, measures like SERVPERF-R, SERVPERF, and SERVQUAL, although they appear on the surface to possess good psychometric properties, will continue to return scores that are uniformly located towards the upper end of the subscales.

In noting the limitations of this research program, it has to be acknowledged that it has not succeeded in its main aim of extending the measurement range of SERVPERF or demonstrating ways by which this could be achieved. The validation of SERVPERF-R in and Australian context has to be regarded as a minor achievement if the derived scale suffers from the same problems as SERVPERF itself. Rasch analysis provides an answer to Parasuraman et al's (1994) plea for appropriate rescaling techniques for upwardly-biased items but it cannot create the range required for adequate discrimi-

nating among customers. Further work is required on the instrument itself for this to occur.

References

Andrich, D. (1982). *An extension of the Rasch model for ratings providing both location and dispersion parameters.* Psychometrika, 47, 105-113.

Andrich, D. Sheridan, B., & Lye, A. (1991). ASCORE: Manual of procedures. Faculty of Education, University of Western Australia.

Babakus, E., & Boller, G.W. (1992). *An empirical assessment of the SERVQUAL scale.* Journal of Business Research, 24, 253-268.

Babakus, E., & Mangold, G.W. (1992). *Adapting the SERVQUAL scale to hospital services: An empirical investigation.* Health Services Research, 26(6), 767-786.

Brown, T.J., Churchill, G.A., & Peter, J.P. (1993). *Research note: Improving the measurement of service quality.* Journal of Retailing, 69, 127-139.

Carman, J.M. (1990). *Consumer perceptions of service quality: An assessment of the SERVQUAL dimensions.* Journal of Retailing, 66 (1), 33-55.

Cronin, J.J., & Tayor, S.A. (1992) *Measuring service quality: A Reexamination and extension.* Journal of Marketing, 56, 55-68.

Finn, D.W., and Lamb, C.R. (1991). *An Evaluation of the SERVQUAL scales in a retail setting.* Advances in Consumer Research, 18, 483-490.

Gagliano, K.B., & Hathcote, J. (1994). *Customer expectations and perceptions of service quality in retail a pparel specialty stores.* Joural of Services Marketing, 8, 60-69.

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

Joreskog, K.G., & Sorbom, D. (1993). New Features in LISREL 8. Chicago: Scientific Software.

Marsh, H. (1996). *Positive and negative global self-esteem: A substantively meaningful distinction of artificors*? Journal of Personality and Social Psychology, 70 (4), 810-819.

Parasuraman, A., Zeithaml, V., & Berry, L. (1994). *Alternative scales for measuring service quality: A comparative assessment based on psychometric and diagnostic criteria.* Journal of Retailing, 70(3), 210-230.

Parasuraman, A., Berry, L.L., & Zeithaml, .A. (1991). *Refinement and reassessment of the SERVQUAL scale.* Journal of Retailing, 67(4), 420-450.

Parasuraman, A., Berry, L., & Zeitham, V.A. (1993). *More on Improving Service Quality Measurement*. Journal of Retailing, 69, 140-147.

Parasuraman, A., Zeithaml, V., & Berry, L. (1988). *SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality.* Journal of Retailing, 64, 12-40.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Education Research.

Spector, P.E. (1997). *When two factors don't reflect t wo constructs: How item characteristics can produce artifactual factors.* Journal of Management, 1997, 23(5), 659-677.

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

Measuring Quality

Table 1
Pattern Matrix for Full SERVPERF Scale Using Exploratory Factor Analysis

Item	Hypothesised Dimension	Factors I	II	III	IV	V
Q1	Tangibles	.10	.74	.01	-.08	.08
Q2	Tangibles	-.03	.67	.11	.21	-.11
Q3	Tangibles	-.14	.34	.15	.27	.28
Q4	Tangibles	.12	.65	-.00	.00	.08
Q5	Reliability	.14	-.13	.07	.36	.29
Q6	Reliability	-.08	.09	-.01	.89	-.05
Q7	Reliability	.40	.12	.10	.56	-.07
Q8	Reliability	.22	-.21	.14	.47	.41
Q9	Reliability	.13	.18	-.07	.08	.37
Q10	Responsiveness	.64	.12	.03	.19	.19
Q11	Responsiveness	-.03	.05	.76	.03	.03
Q12	Responsiveness	.05	-.05	.50	.21	.34
Q13	Responsiveness	-.01	.02	.84	.11	-.11
Q14	Assurance	.56	.02	.33	.13	.11
Q15	Assurance	.46	.24	.44	-.06	.00
Q16	Assurance	.15	.23	.17	.39	-.06
Q17	Assurance	.47	.19	-.14	.12	.28
Q18	Empathy	-.19	.30	.53	-.12	.49
Q19	Empathy	.63	.20	.04	-.03	-.09
Q20	Empathy	.17	.11	.22	-.11	.58
Q21	Empathy	.43	.03	.64	-.13	.07
Q22	Empathy	.50	.15	-.11	.04	.32
		Factor Intercorrelation Matrix				
I		1.00				
II		.35	1.00			
III		.36	.43	1.00		
IV		.34	.23	.33	1.00	
V		.40	.27	.39	.32	1.00

Measuring Quality

Table 2

Means, Standard Deviation, and Reliabilities of SERVPERF-M for Dataset 1

Variables	Items	Mean	SD	Alpha
Tangibles	1, 2, 4	17.02	3.00	.80
Reliability	5, 6, 8	18.84	2.12	.74
Responsiveness	11, 12, 13	18.94	2.12	.74
Assurance	14, 15, 17	19.29	1.99	.82
Empathy	18, 20, 22	18.96	2.41	.82

Note: These item numbers are taken from the position of the items in the 22-item SERVPERF scale.

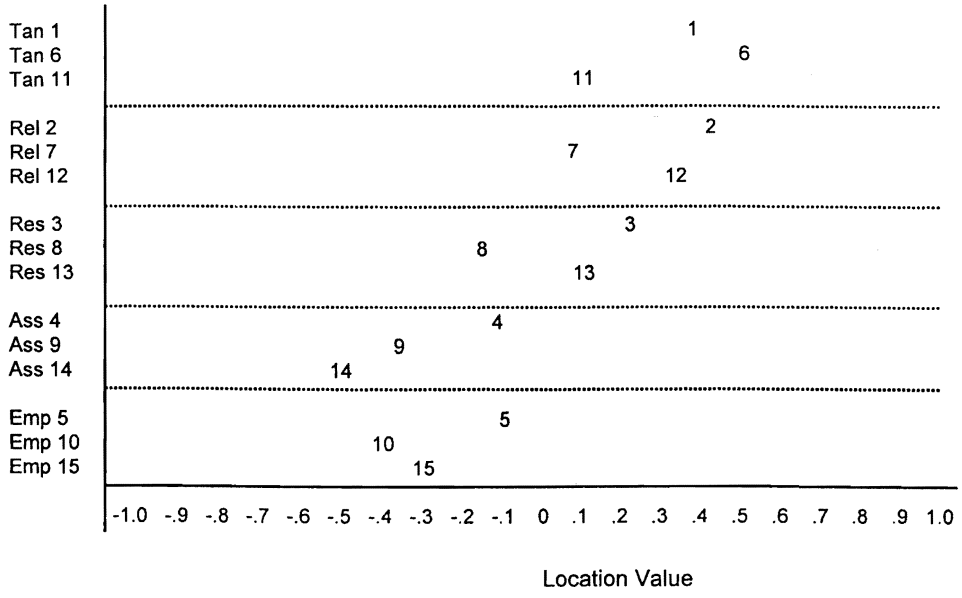
Measuring Quality

Table 3
Range of Domain Covered by SERVPERF-M

Raw Score	Person Freq	Range Logit Score	Person Value Range	Item Value	Range Item	Estimate
89	5	4.9	↑			
88	9	4.21				
87	10	3.73				
86	5	3.38				
85	7	3.09				
84	9	2.85				
83	2	2.63				
82	7	2.43				
81	8	2.25				
80	4	2.09				
79	7	1.94				
78	5	1.79				
77	9	1.66				
76	7	1.54				
75	12	1.42				
74	3	1.31				
73	6	1.21				
72	6	1.11				
71	5	1.01				
70	5	0.93				
69	10	0.84				
68	9	0.77				
67	5	0.69				
66	7	0.62				
65	9	0.56				
64	3	0.49				
63	2	0.43			6	.47
62	2	0.38		↑	2	.39
61	2	0.32			12	.32
60	2	0.27			1	.31
59	3	0.23			3	.16
57	2	0.14			11	.10
56	1	0.09			13	.10
54	1	0.01			7	.06
53	2	-0.03				
52	1	-0.06			5	-.12
50	0	-0.13			4	-.14
49	1	-0.17			8	-.19
47	1	-0.23				
46	1	-0.26				
45	2	-0.30			15	-.30
44	0	-0.33			9	-.32
43	0	-0.36			10	-.36
42	2	-0.39				
39	1	-0.48		↓	14	-.49
36	1	-0.56				
31	1	-0.71				
26	1	-0.87	↓			

Measuring Quality

Figure 1



Appendix A

Customer Satisfaction Survey

To What Extent Do You Agree With the Following Statements About This Company?

		Strongly Disagree						Strongly Agree
1	It has modern looking equipment	1	2	3	4	5	6	7
2	When they promise to do something by a certain time they do it	1	2	3	4	5	6	7
3	They give you prompt service	1	2	3	4	5	6	7
4	You can trust them	1	2	3	4	5	6	7
5	They give you individual attention	1	2	3	4	5	6	7
6	The physical facilities are visually appealing	1	2	3	4	5	6	7
7	When you have problems they show a sincere interest in solving them	1	2	3	4	5	6	7
8	They are always willing to help you	1	2	3	4	5	6	7
9	You can feel safe in your transactions with them	1	2	3	4	5	6	7
10	They give you personal attention	1	2	3	4	5	6	7
11	Materials are visually appealing	1	2	3	4	5	6	7
12	They provide a service at the time they promise to do so	1	2	3	4	5	6	7
13	They are never too busy to respond to your requests	1	2	3	4	5	6	7
14	They have the knowledge to answer your questions	1	2	3	4	5	6	7
15	They understand your specific needs	1	2	3	4	5	6	7
16	They treat you as a very important person	1	2	3	4	5	6	7
17	They do not always have the products you need	1	2	3	4	5	6	7
18	Staff appear to work well together in this store	1	2	3	4	5	6	7
19	They have the very latest goods and styles	1	2	3	4	5	6	7
20	They never object if you want to return goods	1	2	3	4	5	6	7
21	The display of merchandise is less attractive than other stores	1	2	3	4	5	6	7
22	Some staff don't go out of their way to help you	1	2	3	4	5	6	7

Note: Items 16 to 22 added for Study 3 only.

Key:

Tangles:	1, 6, 11	plus 19 and 21 for Study 3
Reliability:	2, 7, 12	plus 17 for Study 3
Responsiveness:	3, 8, 13	plus 20 and 22 for Study 3
Assurance:	4, 9, 14	plus 18 for Study 3
Empathy:	5, 10, 15	plus 16 for Study 3

JOURNAL OF OUTCOME MEASUREMENT®, 4(1), 453-460

Measuring the Capacity of Families to Provide Unpaid Support for a Disabled Family Member: Using BIGSTEPS to Identify Primary and Secondary Dimensions

Kenneth D. Wood

Mark V. Johnston

Kessler Medical Rehabilitation Research and Education Corporation, West Orange, NJ

The Principal Components subroutine of the BIGSTEPS computer program (Rasch analysis) was used to identify primary and secondary dimensions within an item set composed of variables pertaining to family capacity to provide assistance to a family member with disability. Data were obtained through interviews with family caregivers of patients in a major rehabilitation hospital, both during patients' inpatient rehabilitation stay and 3 and 6 months after inpatient discharge. Rasch analysis revealed the primary dimension within these data to be family capacity to provide unpaid instrumental help (18 items) and the strongest secondary dimension to be stressors on the caregivers as assessed during the inpatient stay (5-11 items). A scale consisting of the caregiver stressor items significantly predicted patients' functional gain at 3 months after discharge from inpatient rehabilitation ($r = -.21$) and at 6 months ($r = -.20$), and also the number of days that the patient had spent in a nursing facility as of 6 months after discharge ($r = +.17$). Caregiver stress, burnout, and quality of life at 3 months and 6 months were also significantly predicted. These findings strongly suggest the importance of more definitive research into family stressors that affect long-term patient outcomes.

Requests for reprints should be sent to Kenneth D. Wood, Kessler Medical Rehabilitation Research and Education Corporation, West Orange, New Jersey

Introduction

Rehabilitation hospitals provide inpatient services such as physical therapy, occupational therapy, and cognitive remediation to persons whose disabilities following an injury or illness are too severe or complex for them to return home and receive therapy either there or in an outpatient setting. Patients must not return home while their functional level is lower than can be handled by the help available from families and friends--and paid attendants, if available. Traditionally, assessing the capacity of family and friends to provide help has been done informally and intuitively by social workers and case managers. However, with the current severe pressures from health insurance companies for shorter patient stays, there is increasing risk of patients being sent home before they can safely be managed by the persons available. Objective means for assessing family capacity to provide help are needed, yet at present are not available in the published literature. We have been using the BIGSTEPS program (Wright and Linacre, 1996) to develop a measure of family capacity to provide unpaid help. BIGSTEPS has also identified a caregiver stressor scale based on the residual variation not accounted for by the Rasch-generated model. This secondary scale is the topic of this paper.

Sample

Two hundred and twenty-nine caregivers were interviewed while their disabled family member was an inpatient at a major rehabilitation hospital. Having some family support is an admission requirement to this rehabilitation facility, so practically all persons in the sample had a caregiver who could provide some degree of instrumental support. One hundred ninety-five of these caregivers, plus the former patients, were interviewed by telephone at 3 months post-discharge from the rehabilitation hospital and 191 of these at 6 months in order to obtain outcomes data. Of the 229 family members with disability, 39% were in rehabilitation due to stroke, 20% due to orthopedic problems such as hip and knee fractures or replacements, 14% due to a traumatic brain injury, 13% because of a spinal cord injury, and 14% due to other diagnoses. Patients averaged 61 years old, with a range of 15 to 96 years. Fifty-three percent were female and 47% were male. The ethnic

composition of the sample was similar to that of the hospital's admissions: 71% non-Hispanic caucasians, 20% African American, 5% Hispanic, 2% Asian, and 1% of some other ethnic group. At 3 months post-discharge 17% of the patients were living alone; at 6 months 20% were doing so.

Methods

A survey instrument was developed to assess (1) family capacity to provide instrumental support (i.e., physical help), and (2) factors that affect the capacity of families to provide instrumental support. A pilot study by one of the authors (Johnston, 1983) had previously identified important domains of family capacity to provide instrumental support as well as some good assessment items. Additional domains for inclusion were suggested by literature review (Johnston, Zorowitz and Nash, 1994). Topics covered in the survey instrument are shown in Table 1. Family disfunctionality was assessed with the "affective responsiveness" and "behavior control" subscales of the Family Assessment Device (FAD) (Epstein, Baldwin and Bishop, 1983). Caregiver burnout was assessed with a modification of the Maslach Burnout Inventory (Maslach, Jackson and Leiter, 1996).

Inpatient interviews assessed the family's capacity to provide help as well as demographic and other background information about the patient and family. Questions about sensitive issues that caregivers may prefer to answer in private were asked in a questionnaire that the caregiver filled out while the researcher was absent from the interview room. Examples of potentially sensitive items include: how well the caregiver gets along with the patient, how committed the caregiver is to the patient, how well the caregivers themselves get along, and family dysfunction. Data on patient and caregiver outcomes were collected with telephone interviews at 3 months and 6 months after inpatient discharge. Outcomes assessed include: institutionalization of patient, caregiver stress and burnout, injury of the patient, injury of the caregiver, hospitalization of the patient, hospitalization of the caregiver, and quality of life of patient and caregiver.

All items related to family caregiving, as identified by factor analysis and intuition, were Rasch analyzed using BIGSTEPS. Items with a

Table 1

Major Areas covered in the Survey Instrument

Total Unpaid Help Available (from all family members and friends)

- Difficulty of providing various amounts of helping time.
- Difficulty of performing each of a wide variety of caregiving tasks
- Total hours/day of "on call" help that could be provided
- Total hours/day of "engaged" help that could be provided
- Number of times a week when no one could be with the person

Adequacy of Family Help Available (estimate of extent to which help available will not meet or will exceed help need.)

Information about Individual Caregivers

- Whether lived with the person prior to injury/illness
- Preference for person's residence (home vs nursing home)
- How difficult to live with the person
- How far away living from the person
- Transportation problems in getting to person's residence
- Other obligations (work, family, studies, other)
- Flexibility of schedule
- Prior experience in caring for a person with disability
- Prior medical/attendant training
- Main caregiver's attitude toward hiring paid help for the person
- Disabilities that would affect caregiving
- Endurance and physical strength

Financial Resources of the Person and Family

Quality of Interpersonal Relations between Person and Family/Caregivers, and among Caregivers

Architectural impediments/barriers

Demographics and Conditions Prior to Injury/Illness

- Marital status
- Employment status
- Ethnicity
- Disability prior to injury/illness

MNSQ of 1.5 or greater were deleted from the item set (because of the predominance of noise in these items) through a series of BIGSTEPS runs. The Principal Components Analysis subroutine of BIGSTEPS identified secondary dimensions that account for residual variation not explained by the Rasch model. The predictive validity of the Family Help Available (FHA) scale and one interesting secondary scale was indicated by their correlations with patient and caregiver outcomes at 3 months and 6 months after discharge.

Table 2

Distribution of Rasch-Selected and ranked Family Capacity Items

From most difficult (at top) to least difficult
24-hours a day of unpaid help that requires continuous monitoring
24 hours/day of unpaid help in which the caregiver must be within earshot
Manage a behavior disorder that requires continuous vigilance
Manage a mild behavior disorder
Tube feed
Manage bowel incontinence
Manage urinary incontinence
6 to 8 hours a day of unpaid help
Assist with toileting
3 to 4 hours a day of unpaid help and live with the person
Prepare all meals and feed
Dress and groom
Do all household chores
3 to 4 hours a day of unpaid help
8 hours a week of unpaid help
Check in every so often
Provide transportation within the community
Occasional unpaid help

Analysis and Results

BIGSTEPS identified 18 items that constitute our measure of family capacity to provide unpaid help, which we refer to as the scale of Family Help Available (FHA). These items measure the willingness and ability of family members and friends to provide various amounts of caregiving time and perform various caregiving tasks. These items and their ranking from most difficult to least difficult are shown in Table 2. The FHA has good scaling characteristics: item separation reliability = .99, item separation = 8.34, and person (family) separation = 3.56.

In addition to the FHA dimension, four other dimensions were identified by the Principal Components Analysis subroutine in BIGSTEPS. The strongest of these four dimensions, accounting for 11% of the residual variation, is best described as caregiver stressors. Only this dimension suggests a scale based on items not in the FHA dimension. Table 3 shows the items that load on this factor. The negative end of this factor consists of items that indicate stressful conditions for the caregiver. From these negatively-loaded items we used BIGSTEPS to calculate two stressor measures. One measure consists of the five stron-

Table 3

Factor 1 from principal component analysis of standardized residual correlations

Loading	Measure	Items
.68	-1.20	3 hrs/day unpaid help
.57	-3.40	8 hrs/week unpaid help
.57	-.30	3 hrs/day unpaid help
.49	-5.90	Occasional unpaid help
.48	-.30	Prepare meals and feed
.46	-1.00	Do household chores
.45	4.50	24 hrs/day unpaid help
.42	-1.00	Dress the patient
.40	1.50	6 hrs/day unpaid help
.40	5.50	24 hrs/day unpaid help
.35	3.60	Manage a behavior disorder
.27	2.40	Manage a mild behavior disorder
.22	-3.70	Check in every so often
.21	1.70	Tube feed
.21	-.05	Bathe the patient
-.57	.85	Make ends meet
-.54	1.04	Doesn't fear being overwhelmed
-.46	.10	Caregiver enjoys time with patient
-.45	-.07	Patient doesn't anger caregiver
-.39	.63	Prepared to take patient home now
-.30	.11	Number of possible residences
-.30	.79	Caregivers confident can manage caregiving tasks
-.25	-.38	Caregiver is committed to patient
-.24	.65	Adequacy of help from main caregiver
-.23	-.12	Caregiver's endurance
-.21	.33	Take to appointments

gest stressor items, loading between -.57 and -.39. The second measure consists of these 5 items plus the 6 next-strongest items, loading between -.38 and -.21. The reliability of both scales is good (.93) but person separation is poor.

Predictive validity of the stressor scales. In order to explore the possible practical importance of the two stresssor scales, we calculated their zero-order correlations with selected outcome variables (see Table 4). Both stressor scales significantly predict the following outcomes: number of days that the former patient spends in a nursing home, caregiver burnout, caregiver stress, and caregiver quality

Table 4

Correlations between stressor scales and selected outcomes

OUTCOMES		5-item Stressor Scale	11-item Stressor Scale
Person's quality of life	3 mos	-.09	-.08
	6 mos	-.14	-.12
Person's functional gain	by 3 mos	-.21*	-.14
	by 6 mos	-.20*	-.14
Days in a nursing facility	by 6 mos	+.17*	+.21*
# of ER visits	by 6 mos	+.04	+.06
# of times rehospitalized	by 6 mos	-.03	+.01
Caregiver burnout	3 mos	+.29**	+.28**
	6 mos	+.25**	+.26**
Caregiver stress	3 mos	+.28**	+.30**
	6 mos	+.26**	+.30**
Caregiver's quality of life	3 mos	-.25**	-.29**
	6 mos	-.23**	-.24**
Caregiver injuries	by 6 mos	+.05	+.11

** = $p < .01$
* = $p < .05$

of life. The 5-item stressor scale also predicts the person's functional gain at 3 months and 6 months.

Discussion

In this study the stressor items are shown to form a distinct factor that is not part of the measure of family capacity to provide unpaid help. Because investigating stressors faced by caregivers was not the primary purpose of this study, our selection of stressor items was rudimentary and not systematic. For this reason, the predictive validity demonstrated by these scales should be seen as evidence of the importance of caregiver stress to outcomes for both patients and caregivers. Developing more rigorous measures of caregiver stressors may clearly have practical importance in understanding and predicting patient outcomes.

Conclusion

These findings demonstrate the usefulness of the principal com-

ponents subroutine in BIGSTEPS for identifying secondary dimensions within an item set. By using this new addition to BIGSTEPS, our study has taken initial steps toward developing an objective measure of caregiver stressors, but much more work is needed.

References

Epstein, N. B., Baldwin, L. M., and Bishop, D. S. (1983). The McMaster family assessment device. *Journal of Marital and Family Therapy, 9*, 171-180.

Johnston, M. V. (1983). *The Costs and effectiveness of stroke rehabilitation: Measurement and prediction*. Ph.D. dissertation for the Claremont Graduate School. Claremont, CA.

Johnston, M. V., Zorowitz, R., and Nash, B. (1994). Family help available. *Topics in Geriatric Rehabilitation, 9*(3), 38-53.

Maslach, C., Jackson, S. E., and Leiter, J. P. (1996). *Maslach burnout inventory manual, 3rd edition*. Palo Alto, CA: Consulting Psychologists Press.

Wright, B. D., and Linacre, J. M. (1996). *A user's guide to BIGSTEPS: Rasch model computer program*. Chicago: MESA Press.

JOURNAL OF OUTCOME MEASUREMENT®, 4(1),461-481

The Stability of Health Status Measurement (SF-36) in a Working Population

Jin-Yuan Chern, Ph.D.
Chang Jung University - Tainan, Taiwan

Thomas T. H. Wan, Ph.D.
Virginia Commonwealth University

Michael Pyles, Ph.D.
Virginia Commonwealth University

This study tests the stability of health status measurement (SF-36) in a working population. A total of 4,225 employees from two sectors (one state agency, one private company) enrolled in three health plans at Trigon BlueCross / BlueShield of Virginia. An eight-dimension short-form health survey (SF-36) was first tested on a cross-sectional basis for its validity. Then, a panel study was established to test for the stability of health status instrument over time. Structural equation modeling built on equality constraint conditions was the statistical technique for this study. Data were collected through two-wave mail surveys. Both comprehensive (original eight scales) and parsimonious (revised five scales) models of health status were found fit into the data quite well. Furthermore, the revised parsimonious model was shown highly stable over time. Within a working population aged 18 to 64, people are relatively healthy. Their perception of health issues is reflected mainly on "physical health status," as indicated by physical functionings or role limitations. The high stability of revised health status model warrants the possibility of using a more concise health status instrument for the majority of people in working force.

Key Words. health status, SF-36, structural equation modeling, equality constraint, panel study

---

Requests for reprints should be sent to Thomas T.H. Wan, Ph.D., Department of Health dministration, Virginia Commonwealth University, Medical College of Virginia Campus,P.O. Box 980203, Richmond, VA 23298-0203

Despite the debate over various approaches to controlling for escalating health care costs in the U.S., it is commonly agreed that the reduction of health care costs must not be achieved at the expense of individuals' health and health-related quality of care. Therefore, when making policy decisions, if we can identify the health services that contribute comparatively less to better patient outcomes, we can choose the cost-containment efforts that will achieve the most health benefits at the lowest costs (Kravitz and Greenfield, 1995). In other words, it is essential to examine the relationships among the various types of cost-containment strategies, health care costs, and quality of care. To actualize this purpose, a precise assessment instrument is needed to appropriately measure the health outcomes from health plans that have different cost-containment strategies. Health status is the most direct outcome indicator of health care services.

Furthermore, in capitation payment systems, serious concerns have been raised about risk selection and adverse selection. The possible results of those concerns are restricted access to good health services, especially for those most in need (e.g., chronically ill persons), and possible competitive market failures directly due to cost-premium spiral, or so-called "death spiral" (Price and Mays, 1985; Polzer, 1994). A potential approach to alleviating these concerns is risk adjustment. As maintained in one special report on risk adjustment (Lee and Rogal, 1997), "risk adjustment is a corrective tool designed to reorient the current incentive structure of the insurance market." To date, a variety of risk-adjustment models have been proposed in attempting to capture as much variance in health services expenditures as possible. Among the rest, measures of health status have been demonstrated to have high potential for "predicting" later services utilization (Hornbrook and Goodman, 1995; Lichtenstein and Thomas, 1987; Whitmore, Paul, and Beebe, 1989). It is argued that "any comprehensive effort to model health services' use must consider how people view their own general health and functional state, ..." (Andersen, 1995, p. 3). It is even maintained that without valid health status indicators included as predictors of services utilization, it is impossible to construct an adequate capitation formula that prevents risk selection (Van Vliet and Van deVen, 1992). In other words, a valid and sensitive health status in-

strument is the core of risk-adjustment mechanism in capitation payment systems.

Among different approaches to measuring health status, the use of MOS 36-Item Short-Form Health Survey, or abbreviated as SF-36 (Ware and Sherbourne, 1992), has been gaining more and more popularity for both research and administrative purposes (Hornbrook and Goodman, 1995, 1996). While the validity and reliability of SF-36 have been partially evaluated and proved significant as measures of general health and the health-related quality of life (McHorney, Ware, and Faczek, 1993; McHorney et al., 1994; Ware and Sherbourne, 1992; Ware et al., 1993), those studies were mainly conducted with cross-sectional data. In order to justify the legitimacy of using the SF-36 Health Survey as indicators of health status in risk-adjustment models, the stability of health status instrument over time needs closer scrutiny.

Briefly described, the SF-36 Health Survey instrument has eight scales (or dimensions) of health attributes, each comprising several items (36 in total). Each scale measures one health concept: physical functioning (PF), role limitations due to physical health problems (RP), bodily pain (BP), general health (GH), vitality (VT), social functioning (SF), role limitations due to emotional problems (RE), and mental health (i.e. psychological distress and psychological well-being) (MH). A more detailed description of the items in each scale is provided in the SF-36 Health Survey manual (see Ware et al., 1993). Theoretically, the first three dimensions (i.e. PF, RP, and BP) measure mainly physical health status; the last two (i.e. RE and MH) measure mental health status; and, the other three dimensions (i.e. BP, GH and VT) somewhat cross the two principal components (Ware, Kosinski, and Keller, 1995).

METHODS

DATA AND DATA SOURCES

Data were collected through two waves of mailing survey (conducted at the end of each year, 1994 and 1995, respectively).

Data include person-level demographics, membership, and health status of policyholders who enrolled in Trigon Blue Cross/Blue Shield of Virginia and were employees of two sectors (one state agency and one private company) at the time of survey.

Policyholders were randomly selected from the membership list at Trigon BC/BS. The total number of policyholders to be surveyed in the first wave was 14,331. Among the 8,574 policyholders (59.84 %) who responded to the first wave survey, a total of 5,640 (65.78 %) responded to the second wave survey. In recognition of the restrictive data requirements for a longitudinal study, observations with missing data on any variables (1,385 total, or 24.6 %) were eliminated. Thus, this study is based on a panel of 4,255 policyholders.

DESCRIPTIONS OF STUDY SAMPLE

As of July 1994, the mean age of the study sample was slightly above 45. About 63 percent were males. The majority of the subjects were White/Caucasian (85 %). For most, their highest education was between a high school and a college degree, with a few having less than a high school diploma (6 %), and some with postgraduate education (16 %). Three quarters of the study sample were married. More than two thirds (67.4 %) were enrolled in a Point of Service (POS) health plan, 24.3 percent were in a Preferred Provider Organization (PPO), and 8.3 percent were in traditional fee-for-service (TFFS). The household income fell mainly into categories between $20,000 to $59,999 (62.4 %), with 30.4 percent at more than $60,000 and 7 percent at less than $20,000. More than half (57 %) of the study sample was white-collar workers (executives/managers, professionals, and administrators/clerical staff).

Table 1 lists health status as measured by eight indicators from the two-wave survey. Among these health concepts, five scales (PF, RP, BP, SF, and RE) define health as the absence of limitation or disability. For these scales, a highest score of 100 represents no limitations or disabilities observed. For the other three scales (GH, VT, and MH), a score in the mid-range indicates no limitations or disabilities. A score of 100, in contrast, is achieved only when re-

spondents report positive states and evaluate their health favorably (Ware et al., 1993). Overall, the health status scores of the study group were above the norms for the general U.S. population (see Ware et al., 1993, Table 10.1). All eight dimensions of health status showed a positively skewed distribution, indicating a relatively healthy study sample. Except for physical functioning (PF) and general health perceptions (GH), all other dimensions showed an improved change between the first- and second-wave survey, and all changes were within one point of the scores.

ANALYTICAL APPROACH

The statistical analysis is mainly based on structural equation modeling (SEM) technique (Bollen, 1989; Long, 1983a, 1983b; Jöreskog and Sörbom, 1989; Hayduk, 1987), with the aid of the newly developed software package AMOS (Analysis of MOment Structure) (Arbuckle, 1997).

Briefly speaking, the structural equation modeling attempts to explain the relationships among a set of observed variables (i.e. health concepts) in terms of a generally smaller number of unobserved variables or latent constructs (i.e. health status) through confirmatory factor analysis (CFA). Through CFA, the investigator can assess the reliability and validity of the measurements by examining the theoretical and empirical fit between the measurement model and data (Bollen, 1989; Heck, 1998). If the measurement model is developed with two-wave or repeated measures of the observed indicators, the credibility of the measurement instrument can be further evaluated by inspecting the test-retest results or stability coefficients of the measurement model over time.

In this study, using data from the first-wave health survey (year 1994), the construct of health status is validated in a cross-sectional measurement model (see Figure 1), which also serves as to explore a more parsimonious construct of health status; then, based on equality constraint conditions, the stability of health status instrument over time can be evaluated by incorporating in one model with two waves of health status indicators derived from the preceding step.

After model specification is performed, the assessment of

Table 1. Health Status Scores on SF-36: Eight Dimensions

SF-36 Health Concepts	First Wave (1994) (N = 4,255)		Second Wave (1995) (N = 4,255)		U.S. Norms * (N = 2,474)	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Physical functioning (PF)	87.43	18.78	87.39	19.33	84.15	23.28
Role-physical (RP)	85.50	29.52	85.83	29.74	80.96	34.00
Bodily pain (BP)	75.28	22.40	75.86	22.08	75.15	23.69
General health (GH)	73.36	19.26	73.12	19.31	71.92	20.34
Vitality (VT)	61.48	20.17	61.76	19.97	60.86	20.96
Social functioning (SF)	87.91	19.31	88.07	19.58	83.28	22.69
Role-emotional (RE)	88.34	26.60	89.11	26.50	81.26	33.04
Mental Health (MH)	77.58	15.94	77.77	15.84	74.74	18.05

* Based on general U.S. population (Ware et al., 1993, Table 10.1)

model fit should be undertaken. The criteria for assessment of model fit fall into three groups: examination of the solution, measures of overall fit, and detailed assessment fit (Jöreskog and Sörbom, 1989; Bollen, 1989). In the first step, parameter estimates with the right sign and size, standard errors within reasonable ranges, correlations of parameter estimates, and squared multiple correlations are commonly used to check for the appropriateness of each variable. In the second step, the overall model fit is evaluated to see how well the specified model fits the data. The indicators adopted are chi-square statistic, goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), root mean squared error of approximation (RMSEA) and related significance statistics (P_close value), and Hoelter's critical number (Hoelter, 1983). Unlike the general rules applied for the actual value of chi-square and probability, the chi-square statistic should be as small as possible and probability should be larger than 0.05 in order to reject the null hypothesis that the hypothesized model fit the data statistically. For GFI and AGFI, 0.9 or greater is considered to be a good model fit. The smaller the RMSEA (less than 0.05), the better the model fit. The value of 200 in Hoelter's critical number (CN) with an alpha level at 0.05 is considered to be adequate for a good fitted model. The third step is to determine, and adjust for, the possible sources for the lack of model fit, mainly based on modification indices. The assessment criteria are listed and briefly described in Table 2.

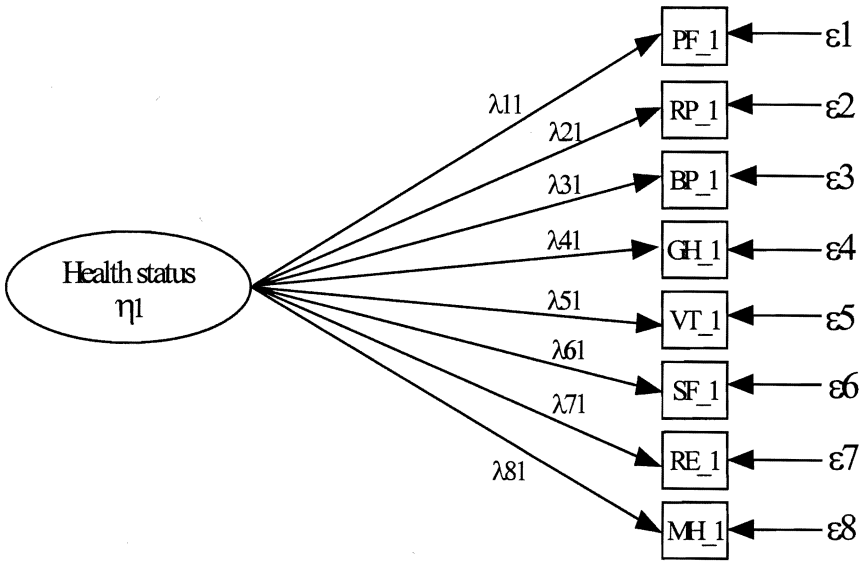
Although a cutoff point of correlation of 0.71 usually is used as the criterion for selection of items (Wan, 1995), a lower threshold of 0.65 is adopted in this study, to retain as many meaningful indicators as possible. In other words, indicators with squared multiple correlations less than 0.4225 are eliminated from the construct, and the consequent (parsimonious) measurement model is used to test for the stability of health status instrument over two points in time.

RESULTS

Validation of Cross-sectional Data for Health Status

Although it is commonly agreed that the construct of health sta-

Figure 1. Path Diagram for Measurement Model of Health Status



Notes:
λ_{i1} (i = 1 to 8) representing factor loadings of health status construct (η_1) on each indicator (i.e. PF_1, RP_1, and so on);
ε_i (i = 1 to 8) representing measurement error of each indicator.

---

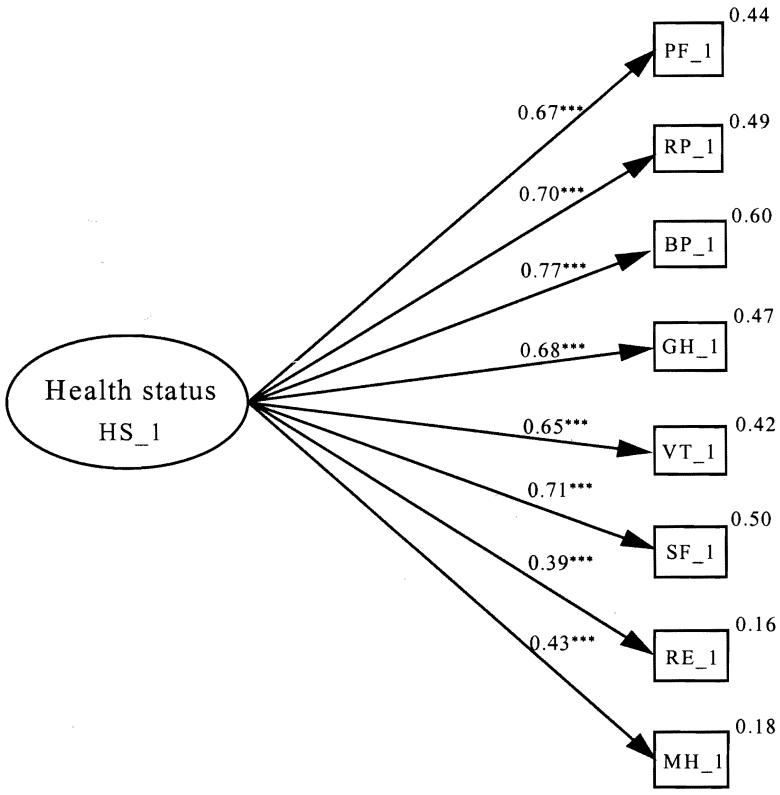
tus is multidimensional, the selection of measurement scales may vary, depending upon the attributes of population measured. Therefore, a measurement model of health status with all eight dimensions was first tested on the cross-sectional data (year 1994). The core results from AMOS are illustrated in Figure 2.

Table 2. Description of Assessment Criteria in Structural Equation Modeling

Statistics	Description	Criterion
Individual Parameters		
C.R. (critical ratio)	C.R. = (Parameter Estimate / Standard Error)	C.R. >= 1.96 (p = 0.05, two-tailed) C.R. >= 2.58 (p = 0.01, two-tailed)) C.R. >= 3.29 (p = 0.001, two-tailed)
SMC (squared multiple correlation)	SMC = (parameter estimate)^2, indicating the strengths of relationships between observed variables and corresponding latent constructs	In measurement models, observed variables with SMC greater than 0.4225 (i.e. estimated R greater than 0.65) are considered significant indicators of latent constructs in this study.
Overall Model Fits		
GFI (goodness-of-fit)	a measure of the amount of variances and covariances jointly accounted for by the model	ranging from 0 (worst) to 1 (best fit)
AGFI (adjusted goodness-of-fit)	a measure of goodness-of-fit while taking into account the degrees of freedom available	ranging from 0 (worst) to 1 (best fit)
RMSEA * (root mean squared error of approximation)	a measure of model adequacy based on population discrepancy in relation to degrees of freedom	RMSEA =< 0.05 (close fit) RMSEA =< 0.08 (acceptable)
P_CLOSE (close fit)	a "p_value" for testing the null hypothesis: H_0: RMSEA =< 0.05	P_CLOSE >= 0.05 (close fit)
C.N. (Hoelter's critical N)	the largest sample size for which one would accept the hypothesis that a model is correct	C.N. >= 200 (satisfactory)

- In contrast, a χ^2 ratio (χ^2 / degrees of freedom) and related p_value measure a test of exact fit. That is, a p_value is used to test for the hypothesis that the population RMSEA is zero; H_0 : RMSEA = 0.

Figure 2. Cross-sectional Measurement Model of Health Status: Initial Model



Overall Model Fit

Goodness-of-fit index	=	1.00
Adjusted goodness-of-fit index	=	1.00
Root mean square error of approximation	=	0.01
P_close value	=	1.00
Critical N	=	5,043

Notes SMCs at upper right corners
Correlated errors not shown
*** 0.001 level of significance

Factor loadings (the numbers shown above each arrow-headed line), comparable to parameter coefficients in general regression models, indicate the relationships (size and direction) between observed variables and their corresponding latent constructs. Squared multiple correlations (those shown at upper right corners), indicating the relative amount of variance in each measurement indicator accounted for by the common construct, measure the strength of a linear relationship between an observed variable and a latent construct (Jöreskog and Sörbom, 1989). Moreover, in structural equation modeling, measurement errors are allowed to correlate with each other to reflect the fact that unexplained variances exist because some variables are missing from the model. By correlating those measurement errors, the model fit can be further improved.

In the initial measurement model of health status, comprising eight SF-36 dimensions, all factor loadings are statistically significant at the 0.001 level, with BP_1 (bodily pain, $\lambda = 0.773$) having the strongest association to health status. Most squared multiple correlations are greater than 0.4225. All correlated measurement errors (θe) are below 0.5, and statistically significant at the 0.001 level. The overall model fit indices show that the model with correlated measurement errors fits the data quite well with goodness-of-fit index (0.999), adjusted goodness-of-fit index (0.996), root mean square error of approximation 0.013, P_close equal to 1, and critical N far beyond 200 (C.N. = 5,043).

As explained earlier, a criterion of factor loadings equal to or greater than 0.65 is used to retain the variables. In the initial measurement model, although all factor loadings are statistically significant, the low squared multiple correlations (factor loadings less than 0.65, or SMC < 0.4225) for VT_1, RE_1, and MH_1 indicate a weaker linear relationship between the latent construct (health status) and those observed variables. A revised measurement model of health status excluding indicators VT, RE, and MH was re-tested for model fit, based on the same data.

As shown in Figure 3, not only are all factor loadings statistically significant at the 0.001 level, but all squared multiple correla-

tions are greater than 0.4225, indicating that at least 42 percent of the variance in each observed variable is captured by the latent construct. GH_1 (general health perception, $l = 0.74$, $p < 0.001$) is the most influential indicator of health status. Furthermore, the overall evaluation of model fit is as good as in the comprehensive initial model with GFI = 1.0, AGFI = 1.0, RMSEA = 0.01, P_close = 0.98, and C.N. =10,016. Therefore, the revised parsimonious measurement model was used to test for its stability over time.

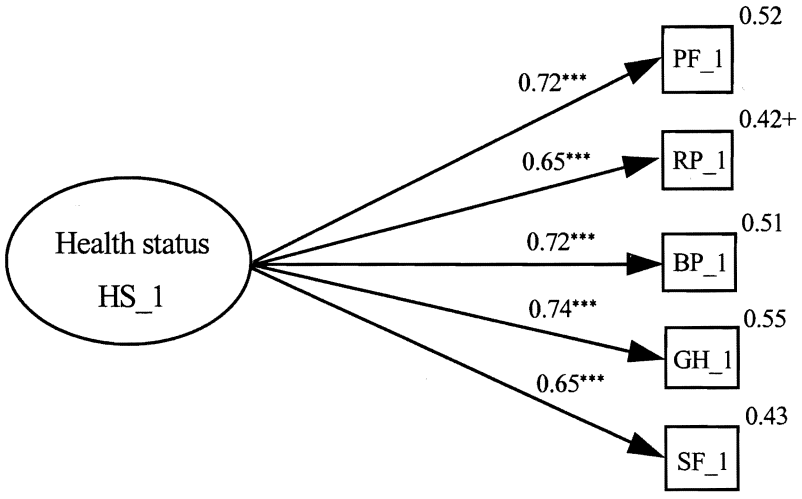
STABILITY OF THE MEASUREMENT MODEL

The validated parsimonious measurement model of health status based on cross-sectional data was used to test for its stability over time by incorporating two waves of health status measures in a panel model. The stability or test-retest reliability coefficient can be generated to show the integrity of the measurement instrument when the same indicators are used to measure the common latent construct such as health status.

For validating the stability of the measurement instrument, equality constraint conditions are assumed for the corresponding factor loadings between each observed variable and the latent construct over two occasions (i.e. $l_{i,t} = l_{i,t+1}$, i = 1 to 5, representing each observed variable, t indicating occasion 1, and t+1 indicating occasion 2). It is very plausible to assume that measurement errors are autocorrelated over time; that is, measurement errors incurred at occasion 1 are probably incurred at occasion 2 (Jöreskog and Sörbom, 1989). Figure 4 illustrates the modeling results.

As expected, all factor loadings are statistically significant at 0.001 level, and all squared multiple correlations are greater than 0.4225, which indicates the appropriateness of using the parsimonious measurement model of health status at different points in time. Most important, the statistically significant high parameter estimate between HS_1 and HS_2 ($\beta_{21} = 0.828$, $p < 0.001$) proves the relatively strong stability of the health status measurement instrument with five scales over time. The autocorrelations are all statistically significant, with GH_1 — GH_2 greater than 0.5. All intra-con-

Figure 3. Cross-sectional Measurement Model of Health Status: Revised Model



Overall Model Fit

Goodness-of-fit index	=	1.00
Adjusted goodness-of-fit index	=	1.00
Root mean square error of approximation	=	0.01
P_close value	=	0.98
Critical N	=	10,018

Notes SMCs at upper right corners
Correlated errors not shown
*** 0.001 level of significance
\+ stronger than 0.4225 but weaker than 0.43

struct measurement errors are far below 0.5. Overall model fit indices demonstrate that the proposed model fits the data extremely well with GFI equal to 1.00, AGFI 0.99, RMSEA 0.03, P_close 1.00, and C.N. 1,360. Therefore, the indicators of health status are found relatively stable over time.

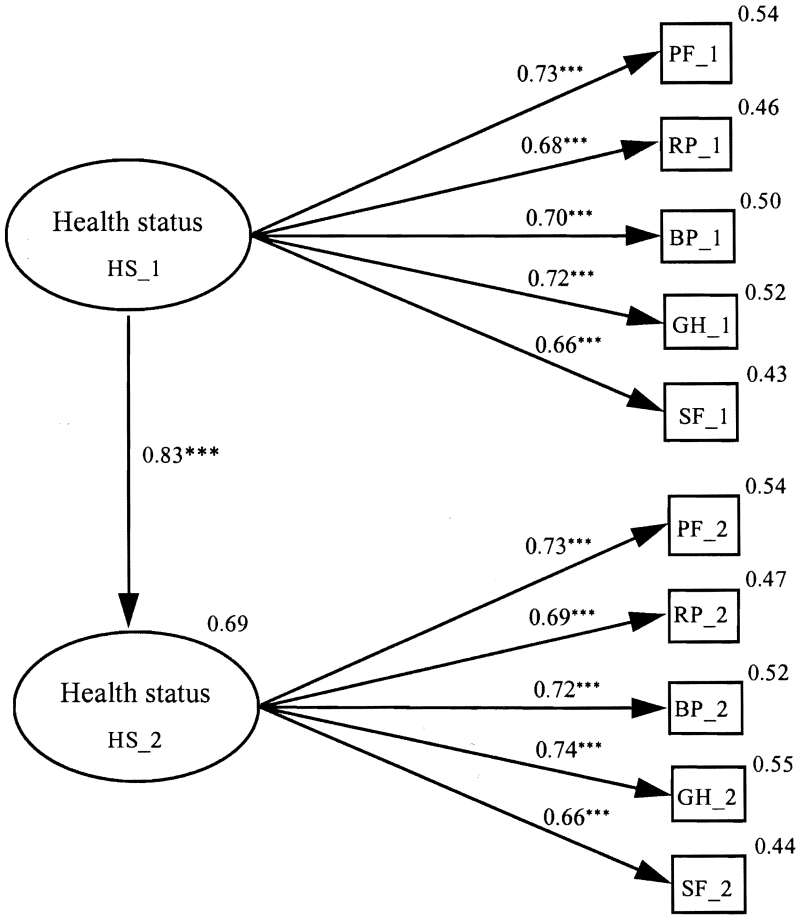
DISCUSSION AND CONCLUSIONS

SUMMARY OF RESULTS

The initial measurement model of eight-scale SF-36 Health Survey instrument fit the data quite well on a cross-sectional basis. However, it failed to demonstrate its validity and reliability longitudinally. Thus, a revised parsimonious five-scale measurement model was developed, and was shown to be well fit cross-sectionally and relatively stable over time, at least over a one-year span. The five scales of health status retained in the measurement model and for the subsequent analyses were physical functioning (PF), role limitations related to physical functioning (RP), bodily pain (BP), general health perception (GH), and social functioning (SF). Three scales excluded were vitality (VT), role limitations related to emotions (RE), and mental health (MH).

At first sight this trimming seems unconvincing, since MH and RE are the two indicators linked mainly to the mental health component, while VT is one somewhat in between physical and mental health components, which, as defined by WHO (World Health Organization, 1976), are the two major dimensions of health status (Ware et al., 1993). A closer scrutiny reveals that health status is not only a multi-dimensional construct, but also one that cannot be decomposed orthogonally. Therefore, although Ware and his colleagues (Ware et al., 1993; Ware, Kosinski, and Keller, 1994) maintained that two principal components were successfully extracted from SF-36 health scales, the two-component solution accounted for just about 81.5 to 84.5 % of variance (as opposed to the original eight-scale instrument), depending on the populations studied. The initial measurement model (with eight scales) for this study sample confirms the oblique relationships among health scales by statistically signifi-

Figure 4. Two-Wave Measurement Model of Health Status—Stability



Overall Model Fit

Goodness-of-fit index	=	1.00
Adjusted goodness-of-fit index	=	0.99
Root mean square error of approximation	=	0.03
P_close value	=	1.00
Critical N	=	1,360

Notes SMCs at up-right corners
Correlated errors not shown
*** 0.001 level of significance

cant correlations of measurement errors. Accordingly, a more parsimonious five-scale measurement model was pursued and found to fit with the data even better. The stability model justified the decision to retain the five scales instead of eight scales as indicators of physical health status. It is possible that the subscales such as vitality, role emotional functional and mental health could be formed into a theoretical construct of mental health status.

Within such a working population aged 18 to 64, people are relatively healthy. Generally speaking, they may not "perceive" health issues until they notice that their work performance or participation in social activities is adversely affected by health problems. In other words, they may view health as one "entity" more related to their responsibilities (work performance) and capabilities (social functioning participation) than to the abstractive personal emotions (mental health, or moods). Therefore, their "physical health status" is reflected mainly by concepts such as RP (role limitations due to physical functioning), BP (bodily pain), GH (general health perception), and RE (role limitations due to emotions), as verified by relatively higher factor loadings on these health scales (see Figure 3). A close examination of the SF-36 Health Survey questionnaires (see Ware et al., 1993) confirms the foregoing inference. In one study comparing the predictive performance of risk-adjustment methods, only PF, RP, BP, and GH scales from the SF-36 Health Survey instrument were retained for analysis, as selected according to the predictive accuracy of indicators (Fowles et al., 1996). Herzlich (1973, cited in Stewart and Ware, 1992, p. 20) has even suggested that "when one is in good health, health is not thought about; health is essentially an 'unawareness' of the body." In summary, within this study group, health status is presumably one construct that can be better defined in terms of roles functioning rather than emotional concerns.

Further, the high stability index ($\beta_{(HS\_1 \rightarrow HS\_2)} = 0.83$) from a longitudinal measurement model suggests a reliable implementation of the health status instrument over time (see Figure 4). Equality constraints proved to be an appropriate assumption, as shown by corresponding standardized regression weights close to each other over time. Statistically significant autocorrelations for measurement

errors are also a necessary condition for such a stability model. In other words, a revised parsimonious measurement model of health status is found relatively stable over time in a working population.

LIMITATIONS

Although the research question in this study is satisfactorily answered, the applications of empirical results are limited in the following aspects.

First, this study involves a relatively "homogeneous" population, in that all sampled observations are for policyholders of three health plans purchased by two employers in one geographic area (i.e. one state). In other words, the results reflect only the phenomenon of health status within a relatively healthy working population. The stability of the five dimensions is influenced by the use of HMO enrollees as the study sample. A study of a diverse population with varying levels of health status may yield different results.

Second, the study sample retained for analysis was screened out through several steps aforementioned, which not only raised concerns over selection bias, but also involved issues about non-response bias and missing values (Rosenbert and Daly, 1993). All these kinds of biases would affect to some extent the generalizability of the study results. Although different strategies may be implemented to take into account the effects of selection bias, non-response bias, and missing values (e.g., pairwise deletion, listwise deletion, imputation, and so on), its implementation does not ensure the representativeness of the sample (Grotzinger, Stuart, and Ahern, 1994). In recognition of the trade-offs between bias and generalizability, this study took no steps to impute values for missing variables, but rather acknowledged caution about generalizability. It is interesting to note that when we performed a separate analysis of the measurement model of health status with missing cases included, similar factor loadings emerged as the case with the listwise deletion of missing cases. In addition, the stability coefficient is slightly lower. However, in the analysis with incomplete data, the statistical program (AMOS) does not display appropriate statistics for testing the goodness of fit. Thus it is difficult to judge the validity of the results.

CONCLUSIONS

The use of general health status instruments has been seen among diverse populations for a variety of purposes, including health policy evaluations, monitoring the health of general populations, and designing systems to monitor and improve health care outcomes (McHorney et al., 1994; Ware, 1995). However, the validity and reliability of health status instrument deserve more attention before its implementation.

This study investigates the stability issue of health status instrument over time, which is often overlooked in other research. A revised measurement model of health status with emphasis on physical health status is proven to be highly stable. This finding provides two suggestions: First, before a health statys instrument can be used (and should be used, as argued by many other researchers aforementioned) for various purposes such as health programs evaluation, risk adjustment for capitation payment systems, and so on, not only does its validity need to be validated for specific groups of population, but also the stability of health status instrument over time needs to be warranted. To put it differently, in a health care market where capitated payment is used, either by insurers to pay health services providers or by ranking the performance of health services organizations, slight changes in the risk adjustment methodology with health status indicators as adjusters would significantly affect the results (Chern, Rossiter, Wan, 2000). Second, conducting a health survey is usually costly and time consuming. Given the results learned from this study, only five indicators are pertinent to the physical health status. Therefore, a shorter-customized health status instrument may be a good alternative, which not only would perform as well as a long and costly health survey, but also would increase the response rate and thus, consequently, would enhance the reliability and validity for the study purpose of health survey.

This study is part of a larger project General Health Study, funded by Trigon Blue Cross/Blue Shield of Virginia, and undertaken by Williamson Institute for Health Studies at Virginia Commonwealth University, Medical College of Virginia campus.

REFERENCES

Andersen, R. 1995. "Revisiting the Behavioral Model and Access to Medical Care: Does it Matter?" *Journal of Health and Social Behavior* 36: 1-10.

Arbuckle, J. L. 1997. *Amos Users' Guide*. Chicago, IL: SPSS Inc.

Bollen, K. 1989. *Structural Equations with Latent Variables*. New York: John Wiley & Sons Inc.

Chern, J. Y., L. F. Rossiter, and T.T.H. Wan. 2000. "Examining the Real Effect of Prior Utilization on Subsequent Utilization." *Research in the Sociology of Health Care* 17: 237-249. In J.J. Kronenfeld (ed.) Health Care Providers, Institutions, and Patients: Changing Patterns of Care Provision and Care Delivery. Stamford, CT: JAI Press.

Fowles, J. B., J. P. Weiner, D. Knutson, E. Fowler, A. M. Tucker, and M. Ireland. 1996. "Taking Health Status into Account When Setting Capitation Rates: A Comparison of Risk-Adjustment Methods." *Journal of the American Medical Association* 276 (16): 1316-1321.

Grotzinger, K. M., B. C. Stuart, F. Aherm. 1994. "Assessment and Control of Nonresponse Bias in a Survey of Medicine Use by the Elderly." *Medical Care* 32 (10): 989-1003.

Hayduk, L. A. 1987. *Structural Equation Modeling with LISREL: Essentials and Advances*. Baltimore, MD: The Johns Hopkins University Press.

Heck, R. H. 1998. "Factor Analysis: Exploratory and Confirmatory Approaches." In G.A. Marcoulides (ed.). *Modern Methods for Business Research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hoelter, J.W. 1983. " The Analysis of Covariance structures: Good-

ness-of-Fit Indices. Sociological Methods and Research, 11: 325-344.

Hornbrook, M. C., and M. J. Goodman. 1995. "Assessing Relative Health Plan Risk with the RAND-36 Health Survey." *Inquiry* 32: 56-74.

Hornbrook, M. C., and M. J. Goodman. 1996. "Chronic Disease, Functional Health Status, and Demographics: A Multi-Dimensional Approach to Risk Adjustment." *Health Services Research* 31 (3): 283-307.

Jöreskog, K., and D. Sörbom. 1989. *LISREL 7: A Guide to the Program and Applications*. (2nd ed.). Chicago, IL: SPSS Inc.

Kravitz, R. L., and S. Greenfield. 1995. "Variations in Resources Utilization among Medical Specialties and Systems of Care." In *Annual Review of Public Health*, edited by G. S. Omenn, J. E. Fielding, and L. B. Lave. 431-445. Palo Alto, CA: Annual Review.

Lee, C., and D. Rogal. 1997. *Risk Adjustment: A Key to Changing Incentives in the Health Insurance Market*. Washington, DC: Alpha Center.

Lichtenstein, R. L., and J. W. Thomas. 1987. "Including a Measure of Health Status in Medicare's Health Maintenance Organization Capitation Formula: Reliability Issue." *Medical Care* 25 (2): 100-110.

Long, J. S. 1983 (a). *Confirmatory Factor Analysis: A Preface to LISREL*. Newbury Park, CA: Sage Publications Inc.

Long, J. S. 1983 (b). *Covariance Structure Models: An Introduction to LISREL*. Newbury Park, CA: Sage Publications Inc.

McHorney, C., J. E. Ware, and A. Faczek. 1993. "The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs." *Medical Care* 31 (3): 247-263.

McHorney, C., J. E. Ware, J. Rachel, and C. Sherbourne. 1994. "The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of Data Quality, Scaling Assumptions, and Reliability across Diverse Patient Groups." *Medical Care* 32 (1): 40-66.

Polzer, K. 1994. "The Role of Risk Adjustment in National Health Reform." *Academic Medicine* 69 (6): 445-451.

Price, J. R., and J. W. Mays. 1985. "Selection and the Competitive Standing of Health Plans in a Multiple-Choice, Multiple-Insurer Market." In *Advances in Health Economics and Health Services Research*, edited by R. M. Scheffler, and L. F. Rossiter. 127-148. Greenwich, CT: JAI Press.

Rosenbert, K. M., and H. B. Daly. 1993. *Foundations of Behavioral Research: A Basic Question Approach*. New York: Harcourt Brace College Publishers.

Stewart, A. L., J. E. Ware. 1992. *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach*. Durham, NC: Duke University Press.

Van de Ven, W., and R. Van Vliet. 1992. "How Can We Prevent Cream Skimming in a Competitive Health Insurance Market? The Great Challenge for the 90's." In *Health Economics Worldwide*, edited by P. Zweifel, and H. E. I. Frech. 23-46. The Netherlands: Kluwer Academic.

Van Vliet, R. and W. Van de Ven. 1992. "Towards a Capitation Formula for Competing Health Insurers: An Empirical Analysis." *Social Science and Medicine* 34 (9): 1035-1048.

Wan, T. T. H. 1995. *Analysis and Evaluation of Health Care Systems: An Integrated Approach to Managerial Decision Making*. Baltimore, MD: Health Professions Press.

Ware, J. E., and C. D. Sherbourne. 1992. "The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual Framework and Item Selection." *Medical Care* 30 (6): 473-481.

Ware, J. E., M. Kosinski, and S. D. Keller. 1994. *SF-36 Physical and Mental Health Summary Scales: A User's Manual*. Boston, MA: The Health Institute, New England Medical Center.

Ware, J. E., M. Kosinski, and S. D. Keller. 1995. *SF-12: How to Score the SF-12 Physical and Mental Summary Scales*. Boston, MA: The Health Institute, New England Medical Center.

Ware, J. E., K. K. Snow, M. Kosinski, and B. Gandek. 1993. *SF-36 Health Survey: Manual & Interpretation Guide*. Boston, MA: Nimrod Press.

Whitmore, R. W., J. E. Paul, D. A. Gibbs, and J. C. Beebe. 1989. "Using Health Indicators in Calculating the AAPCC." In *Advances in Economics and Health Services Research*, edited by R. M. Scheffler, and L. F. Rossiter. 75-109. Greenwich, CT: JAI Press.

World Health Organization. 1976. *Basic Document* (26th ed.). Geneva: World Health Organization.

JOURNAL OF OUTCOME MEASUREMENT®, 4(1),482-490

New Directions in Pediatric Rehabilitation Measurement: The Growing Challenge

Larry H. Ludlow
Boston College

Steven M. Haley
Boston University

The Center on Rehabilitation Effectiveness (CRE) was created in 1998 at Boston University's Sargent College of Health and Rehabilitation Sciences. An important reason for the creation of the Center was demand by purchasers of health services and patients for high quality yet cost-effective rehabilitation programs, particularly with respect to pediatric services. This demand for accountability has created many pressures and challenges for the psychometric community. These challenges include: pediatric rehabilitation assessments that are *conceptually grounded in rehabilitation theory*; new instruments that are *short yet sensitive* enough to detect meaningful disability restrictions and that are sensitive enough to measure meaningful change; and new scales that offer real meaningful comparisons across patients. The purpose of this paper is to explain how the Center for Rehabilitation Effectieness will meet these growing challenges.

Requests for reprints should be sent to Larry H. Ludlow, Ph.D., Boston College - School of Education, 140 Commonwealth Ave., Campion Hall, Chestnut Hill, MA 02467

Our purpose is to introduce the creation of a new center for the study of rehabilitation interventions and innovative models of care, particularly with respect to pediatric services. This center is called the Center on Rehabilitation Effectiveness (CRE). It was created earlier this year at Boston University's Sargent College of Health and Rehabilitation Sciences. Although many research and clinical Centers have focused on the improvement of acute and post-acute medical care, no other academic Center has been created to study exclusively the process and outcomes of pediatric rehabilitation services.

The Center was established to, among other things, coordinate and promote health services and outcomes research in rehabilitation, and to develop and apply new outcome measures in rehabilitation organizations for internal and external comparisons.

The Center considers the study of rehabilitation effectiveness to involve the study of patient co-morbidities and characteristics that influence the rehabilitation process; the formulation and refinement of rehabilitation interventions and delivery systems within an evidence-based practice model; and the examination of the efficiency and outcomes of rehabilitation and post-acute services.

The Growing Challenges

An important reason for the creation of the Center was demand by purchasers of health services and patients for high quality yet cost-effective rehabilitation programs. This demand for accountability has created many pressures and challenges for the psychometric community.

We argue that these challenges are growing and may be viewed as relatively distinct issues.

1. The first challenge is that of ensuring that pediatric rehabilitation assessments have been *conceptually grounded in rehabilitation theory.* Our point is that assessment developers need to recognize that pediatric rehabilitation is a complex process of continuous change for all children yet this change can be characterized and measured when the assessments take into account current rehabilitation theory.

2. A second challenge lies in the need to develop instruments that are *short yet sensitive* enough to detect meaningful disability restrictions and, furthermore, that are sensitive enough to measure meaningful change. The challenge here is to reduce the data gathering burden on clinicians while maintaining a breadth of content that spreads across a continuum sufficiently widely enough that meaningful clinical information can be detected.

3. A third challenge lies in the need to develop scales that are psychometrically appropriate *with respect to linearity* in the sense that assessment scores offer real meaningful comparisons across patients. This challenge is particularly acute when it becomes necessary to make comparisons between patient services and outcomes in rehabilitation programs in different institutions.

This issue sometimes arises in discussions of risk adjusting where the aim may be to look at children from different hospitals where different variables may have different effects on services provided and treatment outcomes. This issue also occurs during discussions of benchmarking of measurements where there is a need to speak the same language across institutions and patients.

Center for Rehabilitation Effectiveness activities

In order to meet these challenges, the Center for Rehabilitation Effectiveness is engaged in five major activities.

The first is the administration of the Rehabilitation Outcomes Management System and DataBase. This database resulted because hospitals and long-term care providers, accredited by the Joint Commission on Accreditation of Healthcare Organizations, have been asked to identify specific performance measures as a strict requirement to meet accreditation requirements. The CRE has been officially approved as a provider of outcomes data that meet the requirements of the accreditation process.

The second activity focuses upon health services research and outcome measurement development. Faculty at Sargent College are actively seeking funds to develop new outcome measures in pediatric rehabilitation and to develop databases for the analysis of rehabilitation services. For example, in addition to the Pediatric Evaluation of Disability Inventory (led by Steve Haley) and the School Function Assessment (led by Wendy Coster), a proposal to develop a post-acute performance measure for children with traumatic injury is pending.

The third activity addresses professional and post-professional training. The Center promotes the development of innovative, outcomes-based professional curricula for the preparation of physical therapists, occupational therapists, speech and language pathologists, and rehabilitation counselors. In addition the Center sponsors post-professional continuing education programs to train experienced practitioners to implement research-based principles for the improvement of pediatric rehabilitation service delivery.

For example, a Summer Institute on Outcomes in Rehabilitation co-sponsored by the Commission on Accreditation of Rehabilitation Facilities (CARF) was held June 25–26 in Boston. The seminar title was "Transforming Outcomes Data into Management Information." This two-day seminar took people through the steps required to complete a data collection project for outcomes management and quality improvement, including:

- identification of resources,
- project design,
- pilot data collection,
- data analysis and interpretation,
- presentations to various audiences, and
- development of quality improvement activities based on research findings.

The fourth activity is in consulting and technical assistance. The Center is active in collaborating with health care organizations to help improve clinical effectiveness, efficiency and the quality of care. For example, we currently work with organizations such as Franciscan

Children's Hospital and Rehabilitation Center, Massachusetts Hospital School, Health Care and Rehabilitation Corporation (HCR), Boston Medical Center, and Pathways Home Care Organization.

Finally, the fifth activity is that of sponsoring faculty and student seminars. The Center facilitates and promotes discussions, seminars, and lectures regarding state-of-the-art methodologies in the development, application, and analysis of outcomes data in pediatric rehabilitation and post-acute healthcare organizations.

The PEDI and SFA Measurement Instruments

The Pediatric Evaluation of Disability Inventory

Thus far the Center has concentrated its assessment development and application efforts primarily upon two instruments—the PEDI and the SFA.

The PEDI project originally began in 1988 through funding from the National Institute on Disability and Rehabilitation Research to develop a state-of-the-art functional assessment instrument for pediatric rehabilitation. It is a comprehensive clinical assessment that samples key functional abilities of children from the ages of six months to 7.5 years. Its intended uses include serving as a discriminative device to determine if functional deficits exist and, if so, the extent and content area of those deficits, an evaluative instrument to monitor individual or group progress in pediatric rehabilitation programs, and an outcome measure for program evaluation of pediatric rehabilitation services or for therapeutic programs in an educational setting.

The PEDI was created from consultations with experts in the fields of pediatrics, pediatric rehabilitation, physical therapy, and occupational therapy. The development of the PEDI was originally based on concepts derived from developmental adaptive tests and models of functional tests used in rehabilitation medicine. It consists of six measurement scales comprised of 197 functional skill/behavior items and 20 caregiver assistance items, and it assesses three major content domains: self-care, mobility, and social function.

Each content domain is assessed by two separate dimensions: functional skills/behaviors, and caregiver assistance. The functional skills items are scored as either 0 for unable to perform the item or 1 for being capable of performing it. The caregiver assistance scales are more exhaustive in their representation of performance capability. The scoring ranges from whether the child is independent (Score = 5) to that level where the child is unable to provide any meaningful assistance during the activity (Score = 0).

The PEDI can be administered by clinicians and educators who are familiar with the child, or by structured interview and parent report. It is designed for use by physical therapists, occupational therapists, rehabilitation nurses, nurse practitioners, speech pathologists, special educators, psychologists, and other professionals who are interested in measuring the functional abilities of young children with disabilities.

Some of the important clinical PEDI research projects have addressed ways to measure and represent meaningful rehabilitation change, and variations in instrument administration due to clinical or school settings.

The School Function Assessment

The *School Function Assessment* (SFA) was designed for the purpose of criterion-referenced skill assessment in the classroom environment. It was developed because the increasing inclusion of students with disabilities in regular education programs necessitated assessments that can identify a student's ability to meet functional as well as academic demands.

There are several adaptive behavior instruments that examine content areas relevant to school function but most of these are norm-referenced instruments designed primarily to define how far below age or grade level a student is performing. Their scores do not provide a clear indicator of which important school-related skills the student has or has not mastered. In addition, most current adaptive behavior assessments do not examine the functional behaviors most affected by physical impairments, nor do they examine whether and to

what extent features of the student's physical and social environment affect functional activity performance.

The SFA addresses these short-comings by providing separate measures of the student's current level of performance on functional activities and the supports needed to perform important functional tasks in the classroom.

Two samples (total n = 607) of elementary school students with disabilities between the ages of 5 and 14 years (mean age = 9.9 years) have been assessed. Data were collected from 120 public school sites across the United States plus Puerto Rico. The students had a variety of disabling conditions with many students having more than one disabling condition.

The content of the SFA examines all major areas of elementary school function with particular attention to areas that are challenging for students with physical or sensory impairments. There are between 10 and 25 items in each scale. Each item is scored on a 4-point rating scale on the basis of the student's typical performance of the particular activity: 1 = Does not/cannot perform the activity, through 4 = Consistent performance (student iniates and completes the activity to a level expected of typical same grade peers).

The SFA is purposely designed so that it can be completed by a variety of school personnel based on instructions provided in the test booklet and without further training. No special training in administration and scoring is provided. Administrators are asked to rely on the description and instructions in the test booklet to understand both the items and the rating scales. Each assessment is typically completed by collaborative effort of two or more persons who work with the student, often a teacher and therapist. These decisions were made to accommodate the constraints on time and resources faced by public elementary schools.

Some of the important school-based SFA research projects have addressed the utility of variable maps to guide definition of outcome groups and to obtain a more in-depth understanding of results. This research has also looked at the utility of classification and regression

tree (CART) analyses. In addition, analyses of change over time have been shown to be more meaningful when the scores are interpreted, using variable maps, in terms of the functional profiles each score represents. In this context, interpretation of results using variable map information on the expected profiles associated with pre and post scores has helped resolve discussions regarding whether statistical differences in scores represent clinically meaningful change.

Conclusion

In meeting the various challenges presented at the start of this article we argue that Item Response Theory (particularly the Rasch model) has enormous potential to help convince assessment developers that it is in their best interests to embrace sound measurement principles when trying to conceptualize and then operationalize functionally different dimensions of assessment. The practical benefits that result from meeting this challenge are particularly evident in the development and application of the Pediatric Evaluation of Disability Inventory and the School Function Assessment scales. These benefits resulted because the scales were explicitly constructed in order to map and measure, in a hierarchical manner, typical pediatric rehabilitation progress.

Through the use of IRT developed scales like the PEDI and SFA it is possible to specify content on the instrument that actually addresses and measures changes in recovery that are meaningful for the rehabilitation needs of patients and their care providers.

Suggested Readings

Coster, W. J., Deeney, T. A., Haltiwanger, J. T., and Haley, S. M. (in press). *School Function Assessment.* San Antonio, TX: The Psychological Corporation/Therapy Builders.

Fisher, A. G., Bryze, K. A., Granger, C. V., Haley, S. M., Hamilton, B. B., Heinemann, A. W., Puderbaugh, J. K., Linacre, J. M., Ludlow, L. H., McCabe, M. A., and Wright, B. D. (1994). Applications of conjoint measurement to the development of functional assessments. *Interna-*

tional Journal of Educational Research, 21(6), 579-593.

Haley, S. M., Coster, W. J., and Ludlow, L. H. (1991). Pediatric functional outcome measures. *Physical Medicine and Rehabilitation Clinics of North America, 2*, 689-723.

Haley, S. M., Coster, W. J., Ludlow, L. H., Haltiwanger, J. T., and Andrellos, P. J. (1992). *Pediatric Evaluation of Disability Inventory (PEDI): Development, standardization and administration manual.* Boston: New England Medical Center and PEDI Research Group.

Haley, S. M., Ludlow, L. H., and Coster, W. J. (1993). Pediatric Evaluation of Disability Inventory: Clinical interpretation of summary scores using Rasch rating scale methodology. *Physical Medicine and Rehabilitation Clinics of North America, 4*, 529-540.

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

Ludlow, L. H., and Haley, S. M. (1996). Effect of context in rating mobility activities in children with disabilities: An assessment using the Pediatric Evaluation of Disability Inventory. *Educational and Psychological Measurement, 56*, 122-129.

JOURNAL OF OUTCOME MEASUREMENT®, 4(1), 491-512

Naturalistic Assessment of Functional Performance in School Settings: Reliability and Validity of the School AMPS Scales

Anne G. Fisher

Colorado State University, Ft. Collins, CO

Kimberly Bryze

University of Illinois at Chicago Early Childhood Research and Intervention Program, Chicago, IL

Bradley T. Atchison

Jamestown, CO

The School Assessment of Motor and Process Skills (School AMPS) is an assessment tool designed to be used by occupational therapists to measure the effectiveness of a student's ability to perform school tasks in naturalistic classroom settings. Rater reliability, internal scale validity, and person response validity of the School AMPS was investigated by examining the goodness-of-fit of raters, motor and process skill items, and students to the many-faceted Rasch model used in the development of the School AMPS. Five of six raters demonstrated acceptable goodness-of-fit ($MnSq \leq 1.4$ and $z < 2$). All 36 motor and process skill items demonstrated acceptable goodness-of-fit. Of the 208 students in the study, 93.7% demonstrated acceptable goodness-of-fit on the School AMPS motor scale and 88.9% demonstrated acceptable goodness-of-fit on the School AMPS process scale. The results of this study support the rater reliability, scale validity, and person response validity for the School AMPS as a tool to be used to evaluate the effectiveness of student performance of school tasks in the classroom setting.

Requests for reprints should be sent to Anne G. Fisher, Colorado State University, Fort Collins, CO.

Introduction

Occupational therapists are concerned with a person's ability to assume roles and to perform those occupations which are relevant to his or her daily life. Occupations are the task performances—purposeful and meaningful activities—a person engages in that support the social and personal roles which define the person. When considering the roles of children and young adults, one's student role is vital. Occupations related to the student role include schoolwork tasks (e.g., writing a narrative entry into a daily journal, using a computer to play an interactive mathematics game, cutting out and gluing construction paper flowers for a springtime art activity), self care (e.g., washing one's hands, donning one's coat), and classroom and student routines (e.g., keeping one's desk organized, turning in assignments on time, passing from class to class) (Bundy, 1995).

In their roles as a related service professional, it is critical, therefore, that occupational therapists use assessments and intervention methods which relate directly to those aspects of the student's school task performances that support his or her student role. The specific emphasis for the therapist is the student's *ability to perform* school tasks rather than the student's acquisition of learning; acquisition of learning is primarily the role of the educator and not the occupational therapist. While the student's success in school may be affected by his or her difficulties performing self care or classroom chores and routines, the student's inability to successfully accomplish schoolwork tasks (e.g., writing, drawing, cutting and pasting) often has a more deleterious effect on the student's performance within the classroom.

The School Assessment of Motor and Process Skills (School AMPS) (Fisher and Bryze, 1997) was developed in response to the need for a valid, reliable, and clinically useful evaluation tool for measuring a student's schoolwork task performance in typical classroom settings. The School AMPS is a naturalistic, observation-based assessment conducted in the context of a student's regular classroom, during his or her typical routine, while the student performs schoolwork tasks assigned by the teacher. Other than the unobtrusive presence of the occupational therapist who observes the student performing

schoolwork tasks, an important feature of the School AMPS is that no disruption of the normal classroom routine occurs during its administration.

The School AMPS is a modified version of the Assessment of Motor and Process Skills (AMPS) (Fisher, 1995, 1997a), which was developed in response to the need for standardized functional assessments which measure the quality of an individual's daily life task performances, specifically personal and instrumental activities of daily living (ADL) (e.g., self care, home maintenance, meal preparation). The AMPS is a naturalistic, observational assessment of occupational performance scored in terms of the efficiency, safety, difficulty, and independence of the universal, *goal-directed* ADL motor and ADL process skill actions which are compiled as part of the enactment of one's occupational performance. Fisher and her colleagues have conducted research which has established the AMPS as a valid and reliable assessment for measuring the functional ability of people across cultures (Bernspång and Fisher, 1995b; Goldman and Fisher, 1997; Goto, Fisher, and Mayberry, 1996; Magalhães, Fisher, Bernspång, and Linacre, 1996), across genders (Duran and Fisher, 1996), across age groups (Dickerson and Fisher, 1993; Fisher, 1997a), with a variety of diagnoses (Bernspång and Fisher, 1995a; Doble, et al., 1997; Doble, et al., 1994; Fisher, 1997a; Kottorp, Bernspång, Fisher, and Bryze, 1995; Pan and Fisher, 1994), and between clinic and home environments (Darragh, Sample, and Fisher, 1998; Nygård, Bernspång, Fisher, and Winbald, 1994; Park, Fisher, and Velozo, 1994). Additional studies have demonstrated the unidimensionality of the AMPS scales (Fisher, 1993, 1994, 1997a) as well as the ability to add new tasks as the need arises (Dickerson and Fisher, 1995; Fisher, 1997b).

Like the AMPS, the School AMPS is an evaluation of occupational performance expressed in terms of the quality (efficiency, safety, difficulty, and independence) of the universal, goal-directed school motor and school process skill actions that are compiled to enact schoolwork task performances. To be consistent with the AMPS, therefore, the School AMPS consists of 16 motor and 20 process skill items (see Table 1). Each of these skill items is scored by the examiner using a 4-point rating scale. A score of 4 reflects the highest (competent) perfor-

mance and a score of 1 indicates the lowest (deficit) performance. For each school motor and school process skill item, specific scoring examples have been developed to parallel the scoring criteria in the AMPS, but using examples that pertain to the student's performance of schoolwork tasks (Fisher and Bryze, 1997).

To date, four categories of school tasks have been developed for the School AMPS—pen/pencil writing tasks, drawing and coloring tasks, cutting and pasting tasks, and computer writing tasks. Each category of tasks includes between four and five specific tasks, for a total of 18 School AMPS tasks (see Appendix). These tasks reflect those most commonly performed in preschool and elementary schools, but are also applicable to older students who have disabilities that affect their schoolwork task performances.

Like the ADL tasks in the AMPS, the schoolwork tasks included in the School AMPS are operationally defined and described only in terms of the essential goal and specified objects or materials that should be used. While defining the tasks is important for standardization and calibration of task difficulty, our experience with the AMPS indicates that the task descriptions do not need to specify precisely the details of how they are to be performed by students. Allowing flexibility in the method and content of the tasks has permitted wide applicability of the tasks across classrooms, students, teachers, and cultures (Atchison, Fisher, and Bryze, 1998, in press; Magalhães, 1995).

A major difference between the AMPS and the School AMPS is that students do not choose which schoolwork tasks they will perform nor do they specify the task constraints (i.e., "task contract" in the AMPS). Rather, what tasks students perform and the specific task criteria are determined by the teacher. Moreover, the teacher-specified task criteria are intended for other students in the classroom as fied task criteria are intended for other students in the classroom aswell as the one evaluated, and the student evaluated is observed as he or she performs the teacher-specified tasks along with his or her classmates as part of their typical classroom routine. Therefore, in order to enable assessment of naturalistic classroom task performance, we have had to develop a collaborative interview process in which the occu-

pational therapist obtains specific information from the teacher about his or her expectations for the student's schoolwork task performance. The information gathered during the teacher interview, conducted prior to observing the student, provides the examiner with clear parameters regarding which tools the teacher deems important for task completion (e.g., colored pencils or crayons, but no markers), the expected degree of interaction between the student and the teacher or other students when asking questions or seeking assistance (e.g., if the student may borrow tools and materials from other students, if the student is required to raise his hand when seeking assistance), and other guidelines for student performance which may affect the scoring of particular items during the assessment process (e.g., if the student is expected to turn in his or her paper to the teacher when finished).

Table 1

Universal, Goal-Directed Skill Actions Included in the AMPS and the School AMPS

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

In its broadest conceptualization, the School AMPS offers a systematic and thorough way of examining the transaction between the student, the task, and the environment, and evaluating the quality of a student's schoolwork task performance, measured at the level of disability and not impairment (Fisher, 1998). The School AMPS offers a vocabulary and new way of thinking about what and how a student does what he or she needs and wants to do given the constraints of the task and the physical and social environment.

The School AMPS was originally developed through the work of Magalhães and Fisher as the focus of Magalhães' (1995) doctoral dissertation. Magalhães examined the validity of a pilot version of the School AMPS in three related studies of students in the United States and Brazil. Magalhães found that the School AMPS item difficulty calibrations remained stable across cultural contexts, School AMPS skill items demonstrated acceptable goodness-of-fit to the many-faceted Rasch (MFR) model for the School AMPS, and that the School AMPS was sensitive enough to measure how a student's classroom environment supports or limits his or her schoolwork task performance.

Atchison, et al. (1998), implemented a subsequent study of the reliability and validity of a new, revised version of the School AMPS (Fisher and Bryze, 1997). Their subjects were 54 students between 3 and 7 years of age ($M = 4.0$, $SD = 0.7$) who were typically-developing or who had educationally-related disabilities (e.g., learning disability, developmental disability, multiple disability). Their results supported acceptable goodness-of-fit of students (person response validity), School AMPS motor and process skill items and tasks (internal scale validity), and the rater (intrarater reliability) to the MFR model for the School AMPS ($MnSq \leq 1.4$, $z < 2$). As there was only one rater, they were unable to assess interrater reliability. Moreover, they found that while there were logical explanations for their misfit, three items—Gathers, Restores, and Initiates—failed to demonstrate acceptable goodness-of-fit to the MFR model for the School AMPS. Finally, their sample was comprised primarily of preschool aged students. Atchison, et al. recommended further research based on data from multiple raters who had scored a heterogeneous sample of students from more diverse age groups. The purpose of this study, therefore, was to examine rater reliability, internal scale validity of the School AMPS tasks and skill items, and person response validity for the School AMPS based on a larger sample of typically-developing and disabled students who had been tested by one of six raters. The specific research questions were as follows:

1. Do trained School AMPS raters demonstrate rater reliability as indicated by acceptable goodness-of-fit to the MFR model of the School?

2. Do the School AMPS scales demonstrate internal scale validity as indicated by acceptable goodness-of-fit of the School AMPS motor and process skill items and tasks to the MFR model for the School AMPS?
3. Do students who are typically-developing and who have educational-related disabilities demonstrate person response validity as indicated by acceptable goodness-of-fit of the students to the MFR model for the School AMPS?

Table 2

Student Demographic Characteristics by Student Group

	Typical $n = 81$	Disabled $n = 103$	Risk $n = 24$
Gender			
Male	40	70	17
Female	41	33	7
Race			
White	76	65	23
Black	2	3	0
Hispanic	2	31	0
Other	1	4	1
Age (years)			
M	6.9	7.5	7.5
SD	2.4	3.6	1.9
Range	3 to 11	3 to 15	5 to 11
No. tasks performed			
1	5	16	14
2	56	74	9
3	19	10	1
4	1	3	0

Methods

Participants

The participants in this study were 208 students whose parents or guardians had given consent for them to be assessed with the School AMPS. The participants included students who were typically-developing (n = 81), students with educationally-related disabilities (n = 103), and students who did not have an educationally-related diagnosis, but whose teachers had expressed concern that the children might be "at risk" for experiencing educational delays or difficulty (n = 24). The demographic characteristics of the students are shown in Table 2.

The diagnoses of the students with disabilities are shown in Table 3. All reported diagnoses were based on the report of the teacher or the occupational therapist who worked with the student.

Table 3

Diagnoses of the Students with Educationally-Related Disabilities

Learning disability (LD)	37
Cognitively impaired/Mental retardation (MR)	21
Developmental delay (DD)	12
Autism	6
Sensory integrative dysfunction/Developmental apraxia (SI)	7
Emotional problems/Attention deficit disorder (Psych)	5
Visually impaired	1
Multiple disabilities	14

Procedures

After obtaining informed consent for participation, the occupational therapist (rater) administered the School AMPS according to standard procedures outlined in the manual (Fisher and Bryze, 1997). The rater first interviewed the student's teacher to determine with what tasks the student experienced problems. As part of the teacher interview, the rater also determined what tasks the teacher planned the students to perform and when, as well as any classroom rules or teacher expectations that would need to be considered during scoring. The raters then observed the students in their regular classrooms performing between one and four School AMPS tasks. When they implemented the actual School AMPS observations, the raters sat to the side of the class and unobtrusively observed the students perform the teacher-specified schoolwork tasks and took notes during the students' performances. Later they scored the students' schoolwork task performances based on the detailed scoring criteria included in the School AMPS manual (Fisher and Bryze, 1997).

Six School AMPS raters scored the participants included in this study. All raters had attended a 5-day AMPS training and calibration course and used the AMPS to rate at least 10 additional people. They then attended 3-day course on the School AMPS. They were calibrated as reliable AMPS raters ($MnSq \leq 1.4$; $z < 2$) (Fisher, 1993, 1994, 1997a).

Data Analysis

The following four facets were considered in the data analysis: (a) the ability of the students, (b) the challenge of the tasks, (c) the difficulty of the skill items, and (d) the severity of the rater. The MFR model for the School AMPS, therefore, was expected to conform to the following assertions: (a) a student is more likely to obtain higher scores on easy skill items and tasks than on hard skill items and tasks, (b) easy skill items and tasks are more likely to be easier for all students than are hard skill items and tasks, (c) the raters are more likely to award higher scores for easy skill items and tasks than for hard skill items and tasks (Atchison, Fisher, and Bryze, 1998; Fisher, 1997a, 1997b; Linacre, 1993). School AMPS raters, skill items, tasks, and students that conform to these expectations will demonstrate goodness-of-fit to the MFR model for the School AMPS.

FACETS (Linacre, 1987-94), a many-faceted Rasch analysis computer program, was used to analyze the School AMPS data and to generate the mean square (*MnSq*) and standardized (z) goodness-of-fit statistics that we used to evaluate the goodness-of-fit of the students and the School AMPS skill items and tasks to the assertions of the MFR model for the School AMPS. Our criteria for acceptable goodness-of-fit were $MnSq \leq 1.4$ and $z < 2$. These values were based on the criteria used to develop the AMPS (Fisher, 1993, 1994, 1995, 1997a) and are more strict than those which can be expected for observation-based assessment tools (Wright and Linacre, 1994). More specifically, our analyses enabled us to consider rater reliability using the rater goodness-of-fit statistics, internal scale validity using the goodness-of-fit statistics for the School AMPS motor and process skill items and School AMPS tasks, and person response validity using the motor and process scale goodness-of-fit statistics for each student assessed with the School AMPS.

Results

Rater Reliability

Five of the six raters demonstrated acceptable goodness-of-fit ($MnSq \leq 1.4$; $z < 2$) to the MFR model for the School AMPS motor and process scales. The same rater failed to demonstrate acceptable goodness-of-fit on

both School AMPS scales (see Table 4). We concluded that five of the six raters scored the students consistent with the assertions of the MFR model for the School AMPS.

Table 4

Rater Severity Calibrations (logits) and Goodness-of-Fit to the MFR Model for the School AMPS

Rater	Severity calibration	*SE*	Infit		Outfit	
			MnSq	*z*	*MnSq*	*z*
Motor scale						
1	0.05	.18	0.9	0	0.8	0
2	-0.03	.04	0.8	-5	0.7	-3
3*	0.26	.12	1.6	6	2.2	5
4	0.10	.29	1.3	1	1.7	1
5	-0.03	.04	1.1	1	1.0	0
6	0.05	.07	1.3	4	1.2	1
Process scale						
1	-0.03	.12	1.0	0	0.9	0
2	-0.01	.02	0.9	-3	0.9	-2
3*	0.09	.08	1.6	7	1.6	7
4	0.13	.20	0.7	-1	0.7	-1
5	0.02	.03	1.1	2	1.1	2
6	0.05	.04	1.0	0	1.0	0

* Rater with failure to demonstrate acceptable goodness-of-fit

Internal Scale Validity

The difficulty calibrations and the goodness-of-fit statistics for the School AMPS motor and process skill items are shown in Table 5 and Table 6, respectively. All of the School AMPS skill items met our criteria for acceptable goodness-of-fit to the MFR model for the School AMPS. In like manner, the challenge calibrations and the goodness-of-fit statistics for the School AMPS tasks are shown in Table 7 and Table 8. All of the School AMPS tasks demonstrated acceptable goodness-of-fit to the MFR model for the School AMPS. Considered together, these results suggest that the School AMPS motor and process skill items and tasks work together to define two unidimensional scales, one for school motor ability and one for school process ability. We concluded, therefore, that the School AMPS scales demonstrate internal scale validity.

Table 5

Motor Skill Difficulty Calibrations (logits) and Goodness-of-Fit to the MFR Model for the School AMPS

	Difficulty calibration	*SE*	Infit		Outfit	
			MnSq	*z*	*MnSq*	*z*
Easier items						
Lifts	2.74	.25	1.1	0	0.6	0
Moves	1.98	.19	1.2	1	1.0	0
Endures	1.96	.19	1.2	1	1.5	1
Transports	1.52	.20	1.1	0	0.7	-1
Reaches	1.40	.13	1.0	0	0.7	-1
Bends	0.94	.11	1.0	0	0.9	0
Walks	0.71	.11	0.8	-1	0.6	-3
Coordinates	0.60	.10	1.1	0	0.9	0
Stabilizes	-0.64	.08	1.2	3	1.2	1
Grips	-0.78	.07	1.0	0	1.1	0
Flows	-1.12	.07	1.1	0	1.1	1
Paces	-1.32	.08	1.2	3	1.2	2
Manipulates	-1.62	.08	0.8	-3	0.8	-2
Aligns	-1.63	.08	1.1	2	1.2	2
Calibrates	-1.88	.08	0.9	-1	1.0	0
Positions	-2.86	.11	0.6	-4	0.6	-5
Harder items						

Person Response Validity

Of the 208 students included in this study, 13 (6.3%) failed to demonstrate acceptable goodness-of-fit on the School AMPS motor scale and 23 (11.1%) failed to demonstrate acceptable goodness-of-fit on the School AMPS process scale (see Table 9). Based on our experience with the AMPS (Fisher, 1997a), we expected approximately 6% of the students to misfit; 5% would be expected to misfit by chance based solely on $z \leq 2$. We concluded, therefore, that the overall rate of misfit was as expected on the School AMPS motor scale, but that the overall rate of misfit was higher than expected on the School AMPS process scale. These results suggest that person response validity is acceptable for the School AMPS motor scale, but may be attenuated for the School AMPS process scale.

Table 6

Process Skill Difficulty Calibrations (logits) and Goodness-of-Fit to the MFR Model for the School AMPS

	Difficulty calibration	*SE*	Infit		Outfit	
			MnSq	*z*	*MnSq*	*z*
Easier items						
Searches/Locates	1.40	.10	1.2	1	1.0	0
Chooses Gathers	1.13	.09	1.3	2	1.1	0
Uses	0.96	.09	1.3	3	1.1	0
Inquires	0.90	.09	1.3	3	1.1	0
Adjusts	0.82	.09	1.4	3	1.4	2
Sequences	0.64	.07	1.1	1	1.0	0
Navigates	0.38	.07	1.0	0	1.1	0
Restores	0.35	.07	1.2	2	1.1	0
Terminates	0.16	.08	1.3	3	1.2	1
Heeds	-0.14	.06	1.0	0	1.0	0
Organizes	-0.23	.07	1.3	4	1.2	1
Paces	-0.24	.07	0.8	-2	0.9	-1
Initiates	-0.26	.06	0.9	-2	0.9	-1
Handles	-0.31	.07	1.2	3	1.2	2
Attends	-0.57	.07	0.8	-4	0.8	-2
Continues	-0.70	.06	1.0	0	1.1	1
Notices/Responds	-0.74	.06	1.1	0	1.1	1
Benefits	-0.98	.07	0.7	-4	0.7	-4
Accommodates	-1.02	.09	0.5	-9	0.5	-8
Harder items	-1.56	.09	0.5	-7	0.6	-6

Discussion

Overall, our results support the rater reliability, scale validity, and person response validity for the School AMPS, and suggest that the School AMPS can be used successfully for evaluation of a student's ability to perform schoolwork tasks within naturalistic classroom contexts. One of the benefits of Rasch measurement methods, however, is that Rasch analyses generate detailed goodness-of-fit statistics that can target potential disruptions to the measurement system. In the case of the School AMPS, these disruptions may be due to a lack of consistency in ratings among raters, failure of a set of items or tasks to define a unidimensional scale, or patterns of response among a subgroup of students that suggest the constructed scale is not valid for testing that subgroup. Our findings that one rater and a higher than expected number of students on the School AMPS process scale failed to demonstrate acceptable goodness-of-fit to the MFR model for the School AMPS signaled to us the need to explore the data further for potential sources of disruption to the measurement system. We examined the data for both the motor and process scales of the School AMPS.

Examination of the data revealed the following potential sources of disruption. First, almost 50% of the misfitting subjects on the School AMPS motor scale and approximately 30% of the misfitting subjects on the School AMPS process scale were scored by Rater 3, the only misfitting rater. This occurred despite the fact that Rater 3 had only scored 6.7% of the total sample. Second, of the students who failed to demonstrate acceptable goodness-of-fit, the majority were students with disabilities (11/13 on the motor scale and 16/23 on the process scale) (see Table 9). Moreover, approximately 80% of the misfitting students were in the lower half of the total sample in school motor or school process ability. The median student ability measure for our sample was 2.70 logits on the School AMPS motor scale and 0.90 logits on the School AMPS process scale.

Our finding that the School AMPS motor and process skill items and tasks demonstrated acceptable goodness-of-fit to the MFR model for the School AMPS suggested that the internal validity of the scales was not the source of the disruption. This conclusion was supported

Table 7

Motor Skill Task Challenge Calibrations (logits) and Goodness-of-Fit to the MFR Model for the School AMPS

	Difficulty calibration	*SE*	Infit		Outfit	
			MnSq	*z*	*MnSq*	*z*
Easier tasks						
Spatial answer (CM-1)	0.30	.06	1.0	0	0.9	0
Short answer (WR-3)	0.24	.07	1.0	0	0.9	0
Scribbling (DC-1)	0.19	.07	1.1	1	1.0	0
Free-coloring (DC-3)	0.15	.16	1.2	1	1.2	0
Cut paste - straight lines (CP-1)	0.05	.08	1.0	0	1.0	0
Composition sentence (WR-4)	-0.06	.06	1.0	0	0.9	-1
Word number copy (WR-2)	-0.11	.07	1.0	0	0.9	-1
Color shapes (DC-2)	-0.12	.10	1.1	1	1.1	0
Cut paste - curved lines (CP-2)	-0.15	.10	1.1	1	1.0	0
Composition paragraph (WR-5)	-0.23	.11	1.0	0	1.6	1
Free-drawing (DC-4)	-0.26	.07	0.9	0	0.9	0
Harder tasks						

by our finding that a detailed examination of the misfitting ratings (1.3% of the total number of ratings, $z \leq 3$) revealed no consistent pattern of misfit across items, tasks, raters, or students (including by diagnosis or age) other than those discussed above.

Table 8

Process Scale Task Challenge Calibrations (logits) and Goodness-of-Fit to the MFR Model for the School AMPS

	Challenge calibration	*SE*	Infit		Outfit	
			MnSq	*z*	*MnSq*	*z*
Easier tasks						
Short answer (WR-3)	0.45	.05	0.9	-2	0.8	-2
Spatial answer (CM-1)	0.23	.04	1.1	3	1.2	2
Word number copy (WR-2)	0.09	.05	1.0	0	1.0	0
Composition sentence (WR-4)	0.08	.04	1.0	0	1.0	0
Scribbling (DC-1)	0.01	.04	1.0	0	1.1	1
Composition paragraph (WR-5)	-0.02	.09	1.2	1	1.0	0
Color shapes (DC-2)	-0.02	.06	1.1	0	1.2	1
Free-drawing (DC-4)	-0.11	.05	0.9	-1	0.9	-1
Cut paste - straight lines (CP-1)	-0.12	.05	1.0	0	1.0	0
Free-coloring (DC-3)	-0.14	.10	1.1	1	1.1	1
Cut paste - curved lines (CP-2)	-0.45	.06	1.0	0	1.1	0
Harder tasks						

While further research on a larger sample of students with disabilities will be needed to clarify the source of disruptions, we felt that the evidence supported the disruption as being associated with students who have disabilities. One possibility is that we are seeing evidence that one or more patterns of diagnostic misfits may emerge from the School AMPS data that may be useful for intervention planning (Fisher, 1993, 1997a).

More specifically, Rasch measurement is a person-centered approach to measurement. If analysis of a student's performance across a set of items reveals that the student unexpectedly passed a hard item or unexpectedly failed an easy item, that student will be identified as having a response that fails to fit the MFR model of the School AMPS. Because we have demonstrated that the School AMPS scales are valid (i.e., the skill items and tasks fit the model) and reliable (i.e., the

Table 9

School AMPS Student Ability Measures and Goodness-of-Fit Statistics for the Misfitting Students

Age (years)	Gender	Group or diagnosis*	Ability measure	*SE*	Infit		Outfit	
					MnSq	*z*	*MnSq*	*z*
Motor scale								
7	F	LD	3.18	.57	1.6	1	4.9	2
8	F	MR	2.91	.47	2.3	3	6.3	4
14	M	MR	2.89	.40	1.7	2	1.8	1
7	F	LD	2.57	.43	1.9	2	1.3	0
13	M	Multiple	2.54	.37	1.4	1	2.7	2
9	F	DD	2.38	.32	1.6	2	1.7	1
5	F	Typical	2.21	.38	1.5	1	1.7	1
4	F	Typical	2.12	.32	1.6	2	1.4	0
4	F	LD	1.86	.41	1.8	2	1.7	1
12	F	DD	1.32	.42	1.8	2	2.2	2
8	F	MR	1.28	.42	1.9	2	2.2	2
6	M	Multiple	0.69	.40	1.8	2	1.7	1
12	M	Multiple	-0.03	.39	1.8	2	2.0	2
Process scale								
7	M	Typical	1.78	.24	1.5	2	1.3	0
9	F	Typical	1.62	.27	1.6	2	2.6	3
6	M	Risk	1.34	.21	1.6	2	2.0	2
8	M	Typical	1.15	.26	1.6	2	1.7	1
8	M	Psych	1.07	.32	1.3	1	3.0	3
12	M	Multiple	0.82	.22	1.2	0	1.8	2
3	F	Typical	0.82	.25	1.6	2	1.7	2
6	M	MR	0.51	.28	2.1	4	2.1	3
10	F	Risk	0.49	.24	1.5	2	1.6	2
13	M	MR	0.41	.24	1.5	2	1.7	2
7	F	LD	0.26	.46	2.9	4	3.0	4
8	F	MR	0.21	.23	1.5	2	1.6	2
3	M	Autism	0.14	.25	1.5	2	1.6	2
3	F	LD	0.08	.25	1.6	2	1.5	1
4	M	Typical	0.08	.23	1.4	1	1.6	2
4	M	LD	-0.08	.27	1.8	3	1.9	3
7	M	Autism	-0.29	.22	1.9	4	2.0	4
12	M	Multiple	-0.29	.50	3.3	5	3.4	4
12	F	DD	-0.66	.27	1.7	2	1.9	3
6	M	Multiple	-0.73	.27	1.6	2	1.4	1
10	M	DD	-1.08	.44	2.2	2	2.2	2
3	M	DD	-1.91	.38	2.2	3	1.7	2
8	F	MR	-2.01	.41	2.3	3	2.0	3

* See Table 3 for specific details regarding diagnoses

calibrations have small *SEs*) we have sufficient basis to be concerned about the validity of the responses of those students that "misfit." When consistent patterns of misfit occur across students with similar disabilities such misfits can become diagnostic indicators (Hambleton, 1989). Diagnostic indicators occur when skill item or task calibrations vary consistently across diagnostic subgroups. For the AMPS, with availability of data for large subgroups, we have been able to identify interesting variations across diagnostic subgroups that have the potential to inform clinical practice and enhance the planning of client-centered interventions, yet which are not of a large enough magnitude to disrupt the measurement system (Fisher, 1997a). A fruitful line of future inquiry with the School AMPS will be to determine if similar diagnostic profiles emerge within the School AMPS data.

Although we could not rule out the possibility that Rater 3 was not scoring the School AMPS items in a manner consistent with both the other raters and the scoring criteria in the School AMPS manual, it is important to note that all of the students rated by Rater 3 were students with disabilities. It is also important to point out that 93.3% of the students in this study demonstrated acceptable goodness-of-fit on the School AMPS motor scale and 89.5% demonstrated acceptable goodness-of-fit on the School AMPS process scale. Moreover, the rate of student misfit met our expectations for the School AMPS motor scale.

Overall, our results support the growing evidence (Atchison, et al. 1998; Magalhães, 1995) that the School AMPS is a valid and reliable tool that can meet the demands for objective assessment of students in school systems. Future research will need to focus on the accrual of more data. In addition to monitoring the items, tasks, and students for goodness-of-fit to the MFR model of the School AMPS, research should focus on the examination of the relationship between school motor and process ability among typically-developing children and the ability of the School AMPS ability measures to differentiate among groups of children (e.g., those who are disabled or "at risk" vs those who are typically developing).

References

Atchison, B. T., Fisher, A. G., and Bryze, K. (1998). Rater reliability and internal scale and person response validity of the School AMPS. *American Journal of Occupational Therapy, 52,* 843-850.

Bernspång, B., and Fisher, A. G. (1995a). Differences between persons with right or left CVA on the Assessment of Motor and Process Skills. *Archives of Physical Medicine and Rehabilitation, 76,* 1144-1151.

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

Bundy, A. C. (1995). Assessment and intervention in school-based practice: Answering questions and minimizing discrepancies. In I.R. McEwen (Ed), *Occupational and physical therapy in educational environments* (pp. 69-88). Hawthorne Press

Darragh, A. R., Sample, P. L., and Fisher, A. G. (1998). Environment effect on functional task performance in adults with acquired brain injuries: Use of the Assessment of Motor and Process Skills. *Archives of Physical Medicine and Rehabilitation, 79,* 418-423.

Dickerson, A. E., and Fisher, A. G. (1995). Culture-relevant functional performance assessment of Hispanic elderly. *Occupational Therapy Journal of Research, 15,* 50-68.

Dickerson, A. E., and Fisher, A. G. (1993). Age differences in functional performance. *American Journal of Occupational Therapy, 47,* 686-692.

Doble, S. E., Fisk, J. D., Fisher, A. G., Ritvo, P. G., and Murray, T. J. (1994). Functional competence of community-dwelling persons with multiple sclerosis using the Assessment of Motor and Process Skills. *Archives of Physical Medicine and Rehabilitation, 75,* 834-851.

Doble, S. E., Fisk, J. D., MacPherson, K. M., Fisher, A. G., and Rockwood, K. (1997). Measuring functional competence in older persons with Alzheimer's disease. *International Psychogeriatrics, 9,* 25-38.

Duran, L., and Fisher, A. G. (1996). Male and female performance of the Assessment of Motor and Process Skills. *Archives of Physical Medicine and Rehabilitation, 77,* 1019-1024.

Fisher, A. G. (1998). Uniting practice and theory in an occupational framework. *American Journal of Occupational Therapy, 52,* 509-521.

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

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

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

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

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

Fisher, A. G., and Bryze, K. (1977). *School AMPS: School version of the Assessment of Motor and Process Skills* (research ed.). Ft. Collins, CO: Three Star Press.

Goldman, S. L., and Fisher, A. G., (1997). Cross-cultural validation of the Assessment of Motor and Process Skills (AMPS). *British Journal of Occupational Therapy, 60,* 77-85.

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

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York: American Council on Education/Macmillan Publishing.

Kottorp, A., Bernspång, B., Fisher, A. G., and Bryze, K. (1995). I A D L ability measured with the AMPS: Relation to two classification systems of mental retardation. *Scandinavian Journal of Occupational Therapy,* 2, 121-128.

Linacre, J. M. (1993). *Many faceted Rasch measurement* (2nd.ed). Chicago: MESA.

Linacre, J. M. (1987-1994). *FACETS: Many faceted Rasch measurement computer program.* Chicago: MESA.

Magalhães, L. C. (1995). Assessing motor and process skills during naturalistic classroom observation: Pilot study. Unpublished doctoral dissertation, University of Illinois at Chicago, Chicago. Magalhães, L. C., Fisher, A. G. Bernspång, B., and Linacre, J. M. (1996). Cross-cultural assessment of functional ability. *Occupational Therapy Journal of Research, 16*, 45-63.

Nygård, L. Bernspång, B., Fisher, A. G., and Winbald, B. (1994). Comparing motor and process ability of persons with suspected dementia in home and clinic settings. *American Journal of Occupational Therapy, 48,* 689-709.

Pan, A. W., and Fisher, A .G. (1994). The assessment of motor and process skills of persons with psychiatric disorders. *American Journal of Occupational Therapy, 48*, 775-780.

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

Wright, B. D., and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

Appendix

School AMPS Tasks

Pen/Pencil Writing Tasks

W-1. Circling and connecting

This task involves using a crayon, marker, pen, or pencil to circle letters, words, or numbers, or to connect figures (e.g., dot-to-dot, connecting matching figures).

W-2. Word or number copying

This task involves using a pencil to copy letters, single words, or numbers (e.g., penmanship practice). Practicing writing one's name is an acceptable alternative. The use of a pen to copy letters, single words, or numbers is an acceptable alternative, provided its use is acceptable to the teacher.

W-3. Short answer (numbers or words)

This task involves using a pencil to fill in short answers (i.e., one to two words) in workbooks or worksheets. Short answers to word or number problems commonly are written in blank spaces that are included in workbooks or on worksheets. Writing short word or number problem answers on blank paper is an acceptable alternative. The use of a pen to write the answers also is an acceptable alternative, provided its use is acceptable to the teacher.

W-4. Composition—one to two sentences

This task involves using a pencil to write one to two short sentences. The sentences may be copied from the blackboard or another paper, or created by the student (free-writing). The use of a pen or marker to write the sentence(s) is an acceptable alternative, provided their use is acceptable to the teacher.

W-5. Composition—paragraphs

This task involves using a pen or pencil to write between one-half to a full page of text that is created by the student. Copying sentences from the blackboard or another paper is not acceptable.

Drawing and Coloring Tasks

DC-1. Scribbling

This task involves free-coloring a blank paper in a manner that results in no recognizable figures or objects (e.g., scribbling). A critical feature of this task is that the student is directed to only put color on the page and not to draw any recognizable figure or to color in a predrawn design (e.g., "Put your favorite colors on this page."). Even if the student does draw a recognizable figure or object , he or she is still scored based on the teacher's directions—Task DC-1.

DC-2. Coloring shapes and spaces

This task involves using crayons, markers, or colored pencils to color (fill-

in) *predrawn* designs or pictures. A critical feature of this task is that the spaces to be colored are predetermined by the predrawn lines (e.g., those included in workbooks or worksheets provided by the teacher). Even if the student only scribbles on the page, he or she is still scored based on the teacher's directions—Task DC-2.

DC-3. Free-coloring

This task involves using *crayons or markers* to draw a simple picture on a blank paper and then color in the essential objects or figures. A critical feature of this task is that the student is expected to draw recognizable objects (e.g., "Draw a picture of your house. Show me the color of your house."). Scribbling is not an acceptable alternative. Therefore, if the teacher directs the students to draw and color in recognizable objects, but the student only scribbles, he or she is still scored based on the teacher's directions—Task DC-3.

DC-4. Free-drawing

This task involves using *colored pens or pencils* to draw a complex picture on a blank paper and then color in the essential objects or figures. Drawing a picture with fine details or elaborations is an acceptable alternative to coloring in the essential objects or figures. A critical feature of this task is that the student is expected to draw a picture with embellishments (e.g., "Draw a picture of you house. Be sure to show me where the windows and the doors are."). Even if the student only scribbles or draws a simple picture, he or she is still scored based on the teacher's directions—Task DC-4.

Cutting and Pasting Tasks

CP-1. Cutting and pasting—straight lines

This task involves cutting along straight lines (predrawn or free-form squares or lines) and pasting the cut pieces onto another piece of paper.

CP-2. Cutting and pasting—curved lines

This task involves cutting along predrawn curved lines (circles, hearts, wavy lives) and pasting the cut pieces onto another piece of paper.

CP-3. Pasting with no cutting

This task involves pasting five or more items (e.g., pieces of paper, cotton balls, noodles) onto a flat surface (e.g., piece of paper, paper plate).

CP-4. Cutting with no pasting

This task involves cutting paper with no requirement to paste (e.g., cutting along the edge of a sheet of paper to make fringe, cutting strips of paper to make bookmakers).

Computer Writing Tasks

CM-1. Simple answer or matching—spatial

This task involves using any input device (e.g., mouse, touch screen) to se-

lect an answer to a problem. Selecting the answer requires that the student choose one answer from among several available static objects; the demand is spatial. The student is not scored on turning on or off the computer or opening the program.

CM-2. Academic computer game—spatial/temporal

This task involves using any input device (e.g., mouse, keyboard, touch screen) to play an academic computer game that involves moving objects. Scoring or obtaining the answer requires that the student time his or her response. The student is not scored on turning on or off the computer or opening the program.

CM-3. Keyboard copying

This task involves using any input device (e.g., keyboard, touch screen) to copy or write letters, words, or short sentences. The student is not scored on turning on or off the computer, opening the program, or printing the document.

CM-4. Word processing

This task involves using any input device (e.g., keyboard, voice activated) to copy or write sentences and paragraphs. The student is expected to open the program, print or save the document (data), and exit the program. Turning on and off the computer is optional.

CM-5. Graphics

This task involves using any input device (e.g., keyboard, mouse, touch screen) to create graphic designs (shapes, tables, graphs, histograms). Adding word text to the created design is an expected part of this task. The student also is expected to open the program, save the document (data), and exit the program. Turning on and off the computer and printing the document are optional.

JOURNAL OF OUTCOME MEASUREMENT®, 4(1),513-523

Pseudolikelihood Estimation of the Rasch Model

Arnold Smit, M.A.
Faculty of Work & Organizational Psychology
De Vrije Universtiteit

Henk Kelderman, Ph.D.
Faculty of Work & Organizational Psychology

An estimation method is proposed for the Rasch model on the basis of the pseudolikelihood theory of Arnold and Strauss (1988). A simulation study was conducted to compare the proposed maximum pseudolikelihood estimates with the well known conditional maximum likelihood and unconditional maximum likelihood estimates for the item parameters of the Rasch model. The results show great similarity between the methods.

Requests for reprints of this article should be sent to Arnold Smit, Department of Work and Organizational Psychology, De Vrije Universiteit, v.d. Boechorststraat 1, 1081 BT, Amsterdam, The Netherlands.

Introduction

Due to its flexibility and sound statistical basis, the loglinear model (Bishop, Fienberg, & Holland, 1975; Haberman, 1978, 1979; Andersen, 1980; Fienberg, 1980; Agresti, 1990) can be applied successfully to a wide variety of statistical problems. Application of the loglinear methodology to the Rasch model (Rasch, 1960) resulted in a flexible framework in which tests for specific assumptions are easily formulated (Tjur, 1982; Kelderman, 1984).

Generalizations to polytomous item responses (Agresti, 1993) and multidimensional traits (Kelderman & Rijkes, 1994) have been developed. The main drawback of this methodology is that the amount of computations involved in estimating the parameters becomes enormous when the number of items, say m, is even moderately large.

Two ways of dealing with this problem are known. First, Mellenbergh & Vijn (1981) have proposed a method for estimating the item parameters of the Rasch model by fitting a logit model to the sumscore x item x itemscore table. The estimates of the item parameters are equal to the unconditional maximum likelihood estimates (UML), see Vijn and Mellenbergh (1982) or Blackwood and Bradley (1989). These estimates are known to be inconsistent. The correction $(m\text{-}1)/(m)$ should be applied, see Wright and Douglas (1977) or Andersen (1980). Secondly, item parameters can be estimated from item pairs, irrespective of all other items, the Minchi procedure of Fisher (1974) or the pseudolikelihood estimation of Linden (1986). The reader interested in pseudolikelihood is referred to Arnold and Strauss (1988), Andrich (1988), Choppin (1968), and Zwinderman (1995).

In the present article we propose to estimate the item parameters in the restscore x item x itemscore table. The proposed method will be called conditional maximum pseudolikelihood because it is a pseudolikelihood method derived from the conditional formulation of the Rasch model.

The Rasch model

Let X_{ij} be a dichotomous random variable representing the outcome of the event that person i responds to item j. The random variable will be scored *1* if person i responds correctly to item j, and *0* if person i gives an incorrect response. In the Rasch model (Rasch, 1960) the probability of a correct response is stated as a function of a persons ability parameter θ_i and an item easiness parameter δ_j. The model for a single item is:

$$P(X_{ij}=1;\theta_i,\delta_j)=\frac{e^{\theta_i+\delta_j}}{1+e^{\theta_i+\delta_j}} \qquad (1)$$

and for the complete response pattern, given local stochastic independence:

$$P(\boldsymbol{X}_i=\boldsymbol{x}_i;\theta_i,\delta)=\prod_{j=1}^{m}\frac{e^{\theta_i+\delta_j}}{1+e^{\theta_i+\delta_j}}=\frac{e^{\theta t_i}}{\prod_{j=1}^{m}1+e^{\theta_i+\delta_j}}e^{\sum_{j=1}^{m}x_{ij}\delta_j} \qquad (2)$$

here $t_i=\sum_{j=1}^{m}x_{ij}$ denotes the sumscore of person i. For readability index i will be suppressed in the sequel.

Because every person in the sample introduces a new θ parameter, standard asymptotics do not apply. For the Rasch model an elegant solution exists for this problem. It is well known that for this model the sumscore is a sufficient statistic for the person parameter. This means that all information on θ is contained in the sumscore, or, stated differently, the test cannot differentiate between persons having the same sumscore. Consequently, when we condition on the sum score, Equation (2) no longer dependends on the person parameter. To see this let S_t denote the set of all response patterns $\boldsymbol{x}$ with sumscore t, $S_t=\{\boldsymbol{x}:\sum_{j=1}^{m}x_j=t\}$, then:

$$P(T=t;\theta,\delta)=\sum_{x\in S_t}\frac{e^{t\theta}}{\prod_{j=1}^{m}1+e^{\theta+\delta_j}}e^{\sum_{j=1}^{m}x_j\delta_j}$$
$$=\frac{e^{t\theta}}{\prod_{j=1}^{m}1+e^{\theta+\delta_j}}\gamma_t(\delta)$$

where $\gamma_t(\delta)=\sum_{x\in S_t}e^{\sum_{j=1}^{m}x_j\delta_j}$ are the well known symmetric basis functions. From this it is easily shown, noting that $P(\boldsymbol{X}=\boldsymbol{x}|T=t)=P(\boldsymbol{X}=\boldsymbol{x})$, that the conditional probability does not depend on the person parameter θ :

$$P(\boldsymbol{X}=\boldsymbol{x};\theta,\delta|T=t)=\frac{P(\boldsymbol{X}=\boldsymbol{x},T=t;\theta,\delta)}{P(T=t;\theta,\delta)}=\frac{P(\boldsymbol{X}=\boldsymbol{x};\theta,\delta)}{P(T=t;\theta,\delta)}$$
$$=\frac{e^{\sum_{j=1}^{m}x_j\delta_j}}{\gamma_t(\delta)}=P(\boldsymbol{X}=\boldsymbol{x};\delta|T=t)$$

Formulating the Rasch model conditionally,

$$P(\boldsymbol{X}=\boldsymbol{x};\delta)=P(\boldsymbol{X}=\boldsymbol{x};\delta|T=t)P(T=t)$$

The model in multiplicative form, introducing the new symbol $\phi_{x_j}^{X_j}=e^{x_j\delta_j}$, yields:

$$P(\boldsymbol{X}=\boldsymbol{x};\delta)=\phi_{x_1}^{X^1}\ldots\phi_{x_m}^{X^m}\phi_t^T \qquad (3)$$

where $\phi_t^T=P(T=t)/\gamma_t(\delta)$. The resulting model is a quasi independence model for the item *1* x ... x item *m* x *t* table. It can be shown that the parameter estimates in this model are equivalent to the CML estimates, see Kelderman (1984) or Tjur (1982). In this article we propose to apply the pseudolikelihood methodology to the Rasch model, the estimation scheme will thus be called conditional maximum

pseudolikelihood (CMPL). The CMPL is computationally more attractive.

Pseudolikelihood estimation of the Rasch Model

Like Strauss and Ikeda (1990) we define the pseudolikelihood here as the product of probabilities of the x_j, with each probability conditional on the rest of the data, and estimate the parameters with standard software by fitting a logit model. Let r_j denote the restscore for item j ($r_j=t-x_j$), and, let $S_{r_j}^1$ and $S_{r_j}^0$ denote the sets of response patterns with restscore r. and itemscore *1* and *0* respectively, so

$S_{r_j}^0=\{x:x_j=1 \wedge r_j=t\}$, and

$S_{r_j}^1=\{x:x_j=1 \wedge r_j=t-1\}$.

Formulating the probability for a response pattern in $S_{r_j}^{x_j}$ in terms of the Rasch model as defined by Equation (4) yields:

Note that

$$\sum_{S_{r_j}^1} \phi_{x_1}^{x^1} \ldots \phi_{x_{j-1}}^{x^{j-1}} \phi_{x_{j+1}}^{x^{j+1}} \ldots \phi_{x_m}^{x^m} = \sum_{S_{r_j}^0} \phi_{x_1}^{x^1} \ldots \phi_{x_{j-1}}^{x^{j-1}} \phi_{x_{j+1}}^{x^{j+1}} \ldots \phi_{x_m}^{x^m},$$

both are summations over response patterns of the remaining items with sumscore r_j. The logit for these probabilities is then:

$$\log\left[\frac{P(X \in S_{r_j}^1)}{P(X \in S_{r_j}^0)}\right] = \delta_j + \lambda_{r_j}^{R_j} \qquad (4)$$

where $\lambda_{r_j}^{R_j}=\log[\phi_{r_j+1}^T/\phi_{r_j}^T]$. This results in a very attractive way to estimate the item parameters of the Rasch model.

The question remains if replacement of the complex term $\log[\phi^T_{r_j+1}/\phi^T_{r_j}]$ by a single parameter $\lambda^{R_j}_{r_j}$ influences the estimates of δ_j .

Comparing different estimation schemes using simulated data

To demonstrate the proposed pseudolikelihood method, a simulation was performed. Four times one hundred datasets were generated with 5, 10, 20 and 40 items respectively, each containing one thousand cases. The data were generated by the following procedure. First, *n* theta values were sampled from the standard normal density and *m* delta values were sampled from the uniform density (domain [*-3,3*]). Then with Equation (1), probabilities were computed for every combination of theta and delta values. Finally, the cells of the resulting *n* x *m* matrix were compared to random numbers from the uniform density, to obtain dichotomous variables.

All datasets were analyzed by the methods described earlier: UML, CML and CMPL. The parameters were identified by fixing the first item parameter and the first sumscore - restscore parameter to zero. The UML and CMPL estimates were computed by logistic regression in S-Plus (GLIM module), and the CML estimates were computed using the LOGIMO program (Kelderman & Steen, 1993), this program uses an estimation scheme that prevents the whole contingency table from being stored (Kelderman, 1992). Extra care schould be taken when zero frequencies occur, these are handled by giving them zero weight in the analyses.

Results

The item parameter estimates were compared with the true values using a simple regression model, where the estimates were regressed on the true values. In the ideal case

an intercept of *0* and a slope of *1* is to be expected. The size of the error variance then gives the empirical variance of the parameter estimates. Before comparing the methods the UML estimates were first multiplied by the factor: $(m\text{-}1) / (m)$.

# items	method	intercept	slope	SD regr.
5	CML	0.0295	1.000	0.144
	CMPL	0.0273	1.000	0.145
	UML	0.0283	1.090	0.190
10	CML	-0.0062	0.999	0.119
	CMPL	-0.0057	0.999	0.119
	UML	-0.0038	1.060	0.139
20	CML	-0.0009	0.998	0.123
	CMPL	-0.0016	1.000	0.124
	UML	-0.0008	1.040	0.122
40	CML	0.0095	0.998	0.129
	CMPL	0.0093	1.020	0.129
	UML	0.0100	1.040	0.127

Table 1 : Summary regression true item parameters on estimated item parameters.

Table 1 reports the results of the regression analyses. The CML, UML and CMPL estimates perform equally well. The most notable difference are the slightly less accurate UML estimates for datasets with five items. In Figure 1 the CMPL estimates are plotted against the CML and UML estimates, aggregated over all items and all datasets.

As can be seen from Figure 1 the estimates are almost identical, although the CMPL estimates seem to be a little bit closer to the CML estimates then the UML estimates. When the parameter estimates are regressed on the parameter estimates the prediction formula are: $\delta_{CMPLE} = 0.0003 + 0.985^* \delta_{CMLE}$, and $\delta_{CMPLE} = 0.0024 + 1.030^* \delta_{UMLE}$.

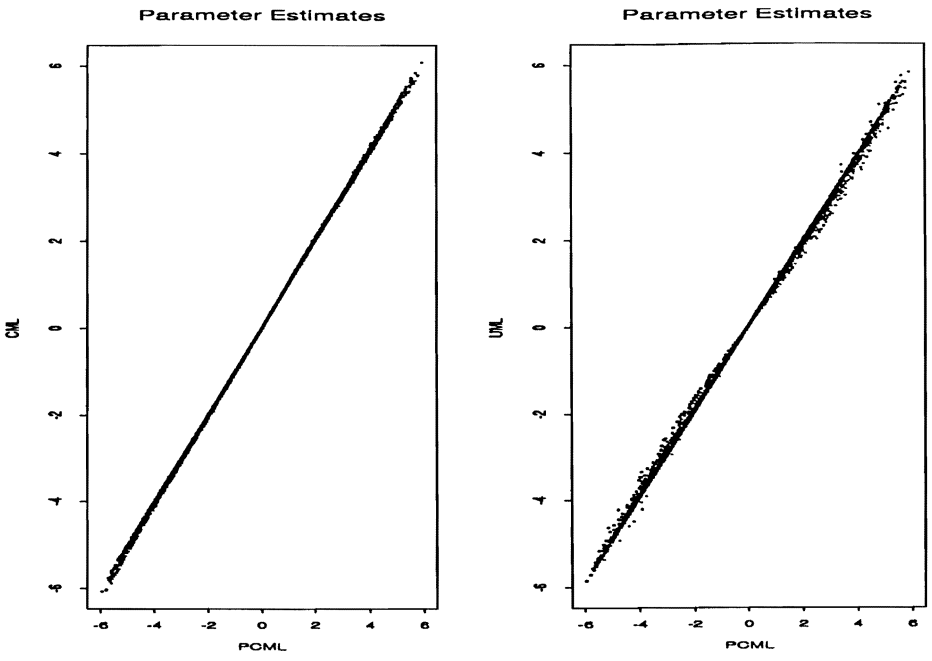


Figure 1: Scatterplot parameter estimates

Discussion

The parameter estimates of the three methods are almost identical. The proposed CMPL method (like the UML method) is computationally more attractive than the CML approach.

The main drawback is that the method can only be used for estimating the parameters. The likelihood ratio statistic

or the Pearson Chi-square statistic of the estimated logit model are not correct. But by reconstructing the expected contingency table (item *1* x ... x item *m* x sumscore) using the parameter estimates it is possible to get the desired likelihood ratio or Pearson Chi-square statistic that would be obtained with the CML method. The proposed pseudolikelihood formulation is less flexible than the loglinear formulation of the Rasch model. The later formulation makes it possible to formulate and test all kinds of interesting hypotheses concerning the test, and the former is merely a convenient way to estimate the item parameters. We conclude by stating that we obtained some promising results with the conditional maximum pseudolikelihood method. Although for the dichotomous Rasch model this method is not directly a serious candidate, generalizations to polytomous items and/or multidimensional latent traits (Kelderman, 1996) rapidly become unmanageable using CML estimation on the full contingency table. With the proposed method it may well be possible to get reasonable estimates. These estimates could in turn be used to obtain fit indices by reconstructing the expected full contingency table.

References

Agresti, A. (1990). *Categorical data analyses*. Wiley: New York.

Agresti, A. (1993). Computing conditional maximum likelihood estimates for generalized rasch models using simple loglinear models with diagonal parameters. *Scandinavian Journal of Statistics, 20*, 63-71.

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. Amsterdam: North-Holland.

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

Arnold, B. C., & Strauss, D. (1988). *Pseudolikelihood estimation* (Tech. Rep. No. 164). Riverside: University of California, Department of Statistics.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivatiate nalyses*. Cambridge: MIT Press.

Blackwood, L. G., & Bradley, L. B. (1989). The equivalence of two methods of parameter estimation for the Rasch model. *Psychometrika, 54*, 751-754.

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

Fienberg, S. E. (1980). *The analyses of cross-classified categorical data*. Cambridge: MIT Press

Fisher, G. H. (1974). *Einfuhrung in die theorie psychologischer tests*. Berne Huber.

Haberman, S. J. (1978). *Analyses of qualitative data* (Vol. 1).New York: Academic Press.

Haberman, S. J. (1979). *Analyses of qualitative data: New developments* (Vol. 2). New York: Academic Press.

Kelderman, H. (1984). Loglinear Rasch model tests, *Psychometrika, 49*, 223-245.

Kelderman, H. (1992).Computing maximum likelihood estimates of Loglinear IRT models from marginal sums. *Psychometrika, 57*, 437-450.

Kelderman, H. (1996). Loglinear multidimensional item response models for polytomously scored items. Pp. 297-304, in *Handbook of Modern Item Response Theory,* (W. J. Van der Linden and R. K. Hambleton eds.). New York: Springer-Verlag.

Kelderman, H. & Rijkes, C. M. P. (1994). Loglinear multidimensional IRT models for polytomously scored items. Psychometrika, 59, 149-176.

Kelderman, H., & Steen, R. (1993). *Logimo: Loglinear IRT modelling [Program Manual]*. iec ProGAMMA: Groningen, The Netherlands.

Mellenbergh, G. J., & Vijn, P. (1981). The Rasch model as a Loglinear model. *Applied Psychological Measurement, 5*, 369-376.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.

Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Society, 85*, 204-212.

Tjur, T. (1982). A connection between Rasch's item analyses model and a Multiplicative Poisson Model. *Scandinavian Journal of Statistics, 20*, 23-30.

Van der Linden, W. J., & Eggen, T. J. H. M. (1986). An empirical bayesian approach to item banking. *Applied Psychological*

Measurement, 10, 345-354.

Vijn, P., & Mellenbergh, G. J. (1982). *The Rasch model versus the Loglinear model.* Révèsz Berichten 42, Amsterdam, Psychologisch Laboratorium.

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1*, 281-294.

Zwinderman, A. H. (1995). Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement, 19*, 369-375.