JOURNAL OF OUTCOME MEASUREMENT®

Volume 2, Number 1 1998

Indexing/Abstracting Services: JOM is currently indexed in the *Current Index to Journals in Education* (ERIC).

Second International Outcome Measurement Conference

Sponsored by
**Rehabilitation Foundation Inc., Institute for Objective Measurement,
MESA Psychometric Laboratory,
and the Journal of Outcome Measurement**

May 15 & 16, 1998
at the International House, University of Chicago

Rehabilitation Foundation Inc., Institute for Objective Measurement, MESA Psychometric Laboratory, and the Journal of Outcome Measurement are sponsoring the Second International Outcome Measurement Conference (IOMC2), an invitational conference on outcome measurement in the health sciences and rehabilitation medicine, at the International House, University of Chicago on Friday and Saturday, May 15 and 16, 1998. The purpose of this conference is to bring together leaders in the outcomes measurement movement in health sciences and rehabilitation medicine in order to facilitate communication among participants and share current research activities. The goal of the conference is to increase the utility, interchangability and communication among instruments used in the measurement of outcomes. The organizers of the conference have invited twenty leading practitioners, including David Cella, Anne Fisher, William Fisher, Norbert Goldfield, Carl Granger, Gunar Grimby, Dennis Hart, Allen Heinemann, Michael Linacre, David McArthur, Mary Segal, Alan Tennant, Luigi Tesio, Craig Velozo, John Ware, and Benjamin Wright, to make presentations on current research in the field of outcome measurement. The presenters are currently doing research on all of the leading outcome scales used in rehabilitation medicine and a number of quality of life scales.

The conference will be structured to allow ample opportunity for participants to discuss the work presented and to formulate strategies for allowing the co-calibration of a variety of outcome measurement systems. This conference is not planned to serve as an introduction to outcome measurement, but as an opportunity for researchers currently developing and using outcome measurement systems to interact and to explore methods of using Rasch measurement models to solve some of the practical problems facing the field of outcome measurement. Although outcome measurement may be applicable to a large number of medical fields, this conference will focus primarily on rehabilitation medicine. Proceedings from the First International Outcome Measurement Conference (IOMC1), held in May, 1996, are available through Hanley and Belfus. To order a volume of the first conference proceedings, please contact them at 215-546-7293.

The conference will be held at the International House on the campus of the University of Chicago (1414 East 59th Street, Chicago, IL 60637 Phone: 773-753-2270). The registration fee of $50 includes a course notebook as well as two buffet breakfasts, two box lunches and four breaks. There is an optional dinner planned for the evening of May 15 at a nearby (walking distance) Italian restaurant. Accommodations are available at the International House or at the Ramada Inn, located in Hyde Park at 49th Street and Lake Shore Drive. The International House has single dormitory style rooms available for $36.00 per night. Call I-House directly to make reservations. The Ramada is offering conference participants a special overnight rate of $89/single, $94 double. Please make reservations directly with the hotel by April 1, 1998. Please inform reservations that you are attending the RFI conference. Ask about the courtesy shuttle to the University of Chicago. Airport transportation is available on C&W Airport Limousine (773-493-2700) for $14 from O'Hare and $11 from Midway to both I-House and the Ramada Inn. Service is also available through Fidelity Limo (312-618-3130).

Registration for the conference is limited to 115 persons, so be sure to register early to reserve a place at the conference. The registration deadline is April 15, 1998. Simply complete the enclosed registration form, include your registration fee and return it to the address listed below. Please feel free to copy this letter and the registration form and distribute it to any of your colleagues who might be interested in attending.

Tentative Schedule:
Friday May 15, 1998

7:30 - 8:30 am	Registration/Continental Breakfast
8:30- 12:00 pm	Presentations and Discussion
12:00 - 1:00 pm	Lunch (provided)
1:00 - 5:00 pm	Presentations and Discussion
6:00 - 8:00 pm	Dinner at Piccolo Mondo

Saturday May 16, 1998

8:00 - 8:30 am	Continental Breakfast
8:30- 12:00 pm	Presentations and Discussion
12:00 - 1:00 pm	Lunch (provided)
1:00 - 3:00 pm	New Research Applications
3:00 - 4:30 pm	Planning Session
4:30 pm	Evaluation/Adjourn

For further information on the conference please contact:
Richard M. Smith
Rehabilitation Foundation Inc.
P.O. Box 675
Wheaton, IL 60189

Voice: 630-462-4102
Fax: 630-462-4547
E-mail: jomea@rfi.org

JOURNAL OF OUTCOME MEASUREMENT®, *2*(1), 4 - 24

The Dimensionality and Validity of the Older Americans Resources and Services (OARS) Activities of Daily Living (ADL) Scale

Susan E. Doble
Dalhousie University

Anne G. Fisher
Colorado State University

The psychometric properties of the OARS ADL scale, comprised of seven physical activities of daily living (PADL) and seven instrumental activities of daily living (IADL) items, were examined using a Rasch measurement approach. Two of the PADL items failed to demonstrate acceptable goodness-of-fit with the measurement model but the remaining 12 items could be combined into a single measure of ADL ability. Although the OARS ADL scale was designed to identify those community-dwelling elderly who need supports and services to continue to live in the community, the scale items were found to be poorly targeted to community-dwelling elderly since almost half of our sample received maximal scores. Rasch analysis identified how we might improve the sensitivity of the OARS ADL scale but its utility in outcome and longitudinal studies remains questionable.

Requests for reprints should be sent to Susan E. Doble, School of Occupational Therapy, Dalhousie University, Halifax, Nova Scotia B3H 3J5, Canada.

Health care providers for the elderly need to be able to identify accurately those persons who experience difficulty when performing activities of daily living (ADL). The ability to competently perform personal activities of daily living (PADL) and instrumental activities of daily living (IADL) has been identified as a key determinant of whether someone can live in the community independently or needs in-home supports and services (Greiner, Snowdon, & Greiner, 1996; Pfeiffer, McClelland, & Lawson, 1989; Rubenstein, Schairer, Wieland, & Kane, 1984; Spector, Katz, Murphy, & Fulton, 1987). Dependency in PADL and IADL has been related to poor quality of life, increased risk of nursing home placement, and increased risk of death (Branch & Ku, 1989; Gillen, Spore, More, & Freiberger, 1996; Greiner et al., 1996; Manton, 1988; Spector et al., 1987; Wolinsky, Callahan, Fitzgerald, & Johnson, 1993).

Information elicited from ADL assessments influences many decisions about the lives of older adults.In some diagnostic situations, such as the assessment of age associated cognitive decline and dementia, the ability to determine whether an individual is able to perform ADL or not is of central importance (American Psychiatric Association, 1994). In fact, the distinction between age associated cognitive decline and dementia is made on the basis of whether an individual's ADL functioning is impaired or not. If ADL assessments are unable to detect deficits in individuals' ADL functioning, incorrect diagnoses will be made. In addition, decisions related to the need for supports and services such as home care, and for long-term care placement are often based on estimates of the person's ability to perform ADL independently. Poon (1994) has also argued that interventions, including behavioral and pharmacological interventions, should be judged effective only if they enable older adults to participate in everyday tasks more effectively. Thus, the need for valid and reliable measures of the ability of older adults to perform PADL and IADL tasks is paramount.

Most commonly, older adults' ability to perform everyday occupations is assessed using self- and proxy-based rating scales. Despite their widespread use, self- and proxy-based assessments of PADL and IADL abilities are plagued by two major problems. First, most ADL assessments evaluate PADL and IADL separately using two different scales which usually contain a small number of items. Since ceiling and floor effects are common, our ability to measure changes in individuals' abilities is reduced. Error is increased by ceiling and floor effects, and consequently, the reliability of the scales is compromised. Although Suurmeijer and colleagues (1994) found that the unidimensionality of the PADL and IADL items was compromised when the items were combined into a single scale, several other studies support the idea of combining PADL and IADL items into a single ADL scale (Finch, Kane, & Philp, 1995; Kempen & Suurmeijer, 1990; Siu, Reuben, & Hays, 1990; Silverstein, Fisher, Kilgore, Harley, & Harvey, 1992; Spector et al., 1987). Generally, items that measure IADL task performance are more difficult than PADL items but there is evidence that some items from the two scales overlap (Finch et al., 1995; Fillenbaum, 1988; Kempen & Suurmeijer, 1990; Siu et al., 1990; Silverstein et al., 1992; Spector et al., 1987; Suurmeijer et al., 1994). Most studies in which the hierarchical ordering of PADL and IADL items have been examined have found that eating and grooming are among the easiest tasks whereas heavy housework and meal preparation are among the hardest tasks. Unfortunately, only a limited number of PADL and IADL were examined in each study. All items which typically comprise PADL and IADL scales should be examined together in order to determine whether they represent the same underlying construct or not.

A second problem with self- and proxy-based assessments is the common practice of summing ordinally rated items to generate a summary score. It has been strongly argued that when ordinally rated items which vary in difficulty are summed, the properties of true measurement are compromised (Fisher, 1993; Merbitz, Morris, & Grip, 1989; Wright & Linacre, 1989). When scales are hierarchically structured using Guttman scaling procedures, the problems experienced when trying to interpret summary scores derived from ordinal ratings are reduced. Guttman scaling procedures have been used to determine if PADL and IADL items can be hierarchically arranged on a linear continuum (Fillenbaum, 1985; Katz, Ford, Moskowitz, Jackson, & Jaffe, 1963; Kempen & Suurmeijer, 1990; Spector et al., 1987). However, as a deterministic model, Guttman scaling requires that the data maintain an absolutely rigid hierarchical structure with items ordered from most to least difficult (Guttman, 1950). More specifically, Guttman scaling is based on the expectation that persons will pass *all* items that are easier than their ability level and will fail *all* items that are more difficult. To ensure that a scale conforms to this expectation of a clear-cut pass/fail point for each person, the differences between item difficulties must be large (Fisher & Fisher, 1993). Thus, the sensitivity of such scales to small changes in functioning within individuals over time or to small differences between individuals is dramatically reduced (Finch et al., 1995). Additionally, the rigid hierarchies expected by Guttman scales have rarely been borne out in either social or behavioral research (Siu et al., 1990; Wilson, 1989). Eating may be easier than bathing for most people but we cannot definitively state that it will *always* be easier for *all* people.

Unlike Guttman scales, Rasch models are probabilistic. They are based on the assertions that persons of a given ability level are *more likely* to pass items that are easier than their ability level and are *more likely* to fail items that are more difficult. Item difficulties are not based on the clinical judgements of experts. Instead, persons who take the test provide us with the empirical data needed to inform us about which items are experienced as more difficult and which are experienced as relatively easy. Item difficulties are estimated on the basis of the probability of persons being able to pass each item. The measure of the difficulty of each item is expressed in logits (log-odd probability units) (Wright & Stone, 1979). As items become easier, the probability of passing items increases to 1.0; as items become more difficult, the probability of passing decreases to zero (Hambleton, 1989; Wright & Stone, 1979). A person's ability level is determined as that point where his or her probability of passing items that match his or her ability is 0.5 (Fisher, 1993; Hambleton, 1989; Wright & Stone, 1979).

Statistics generated through the Rasch analysis can be used to tell us how well responses to test items fit the expectations of the Rasch measurement model (Wright & Masters, 1982). Item separation indices tell us if the items comprising a test are spread out along a linear continuum such that they define distinct levels of difficulty. Rasch analysis also generates a calibrated difficulty level for each test item, a mean square residual (*MnSq*)(with an expected value of 1.0), and a standardized goodness-of-fit statistic (z) (with an expected value of .0). Item difficulty calibrations inform us where on the scale's linear continuum of difficulty each of the items are located. Harder items are expected to be located at one end of the linear continuum and easier items are expected to be located at the opposite end. Items should be spread along the linear continuum without obvious gaps. If the test is appropriately targeted for the ability of the sample being tested, the person ability measures of the subjects should be located within the boundaries of the hardest and easiest items. Also, we can determine if each item contributes to the measurement of the same underlying construct by examining each item's *MnSq* and standardized goodness-of-fit statistics. When an item's *MnSq* is greater than 1.4 and the z greater or equal to 2.0, it indicates that the item is not related to the rest of the items comprising the scale either on a conceptual or practical level.

Rasch measurement procedures also enable us to examine how well the responses of persons fit the expectations of the measurement model (Wright & Masters, 1982). Person separation statistics tell us whether the test separates the subjects comprising a sample into distinct levels of ability. Greater separation suggests increased sensitivity. When other information about the persons taking the test is available, we can examine the placement of the subjects on the linear continuum of ability and determine if their placement is reasonable. Lastly, the validity of each person's ability measure can also be examined using *MnSq* and *z* statistics. The Rasch measurement model asserts that the easier an item is, the greater the probability that an individual will pass it, and the more able a person is, the greater his or her probability of successfully passing an item (Wright & Stone, 1979). Therefore, high goodness-of-fit statistics indicate that a person's pattern of response failed to fit the expectations of the measurement model (Linacre & Wright, 1994a; Wright & Masters, 1982; Wright & Stone, 1979). For example, when an able person fails easy items and when a less able person passes hard items, both persons' responses are unexpected.

One self- or proxy-based assessment of everyday functioning that is used extensively in gerontology research is the Older Adults Resources and Services (OARS) Activities of Daily Living (ADL) scale (Fillenbaum, 1985, 1988). Part of a larger assessment, the OARS ADL scale is often used as an independent assessment of how much assistance individuals need to perform those tasks which enable them to live in the community independently. Like many other self- or proxy-based ADL assessments, it consists of a PADL and an IADL scale. Both of these seven-item scales share a common rating system to determine individuals' ability to perform each of the items independently.

Fillenbaum (1988) proposed no clear theoretical framework although the inclusion of both a PADL and IADL scale reflects Lawton and Brody's (1969) conceptualization that IADL tasks are more complex than PADL tasks. Whereas PADL tasks are largely motoric behaviors which involve the care of one's own body, IADL tasks are performed to ensure that one can be self-reliant in one's own environment. Items for the OARS ADL scale were selected by reviewing items which had been included in other published assessments of PADL and IADL (e.g., Lawton and Brody's (1969) Physical Self-Maintenance and Instrumental Activities of Daily Living scales) and by adding new items which were assumed to reflect a person's ability to live in the community. When the OARS IADL scale was examined to determine if it could be characterized as a Guttman scale, Fillenbaum (1988) reported that a random sample of 1,609 older adults in one American city found housework to be the hardest item. Items assessing the ability to travel, shop, prepare meals, and manage one's own money followed. However, the hierarchical ordering of the items was not stable. Instead, the ordering of some items varied slightly when other samples of elderly adults were assessed.

The purpose of this study was to examine the psychometric properties of the OARS ADL scale using a Rasch measurement approach. First, the extent to which the OARS ADL scale items fit the expectations of the Rasch measurement model was determined. Specific questions about whether the 14 items worked together to define a single variable included: did they define an identifiable linear continuum of increasing difficulty, were the items sufficiently spread out along the linear continuum to define distinct levels, and were the items ordered logically along the linear continuum? Second, the extent to which the OARS ADL scale separated persons along the linear continuum of ADL ability was examined. More specifically, we determined whether the items separated the subjects into distinct ability levels, whether the hierarchical placement of the subjects along the linear continuum of ability was reasonable, and whether at least 95% of the subjects' ability measures were valid (i.e., fit the expectations of the measurement model).

METHODS

The sample of convenience included 372 community-dwelling elderly adults who resided in either Canada or the United States. Canadian subjects were recruited from the Nova Scotia sample of the Canadian Study of Health and Aging (*n*=93), a national epidemiological study of the prevalence of dementia, and from outpatients of geriatric clinics of the Queen Elizabeth II Health Sciences Centre, Halifax, Nova Scotia (*n*=74). The American sample (*n*=203) was recruited as part of a National Institutes of Health, National Institute on Aging study. All of the subjects were at least 60 years old (*M*=74.1 years, *SD*=7.6 years). The majority of the subjects were female (63%). Subjects either lived alone (45.2%), with their spouses (42.2%), with other family or friends (11.5%), or with paid, live-in assistants (1.1%). The majority of the subjects (63.2%) reported having at least one medical condition that affected their everyday functioning; diagnostic information was confirmed through review of clinic data. Twelve percent of the subjects reported having systemic conditions (i.e., cardiovascular, respiratory, gastrointestinal or kidney disease, diabetes, and cancer); 11.5% reported having musculoskeletal conditions (i.e., rheumatoid arthritis, osteoarthritis, hip fracture or replacement, back/neck pain, upper or lower limb injury including amputations, and ankylosing spondylitis), 3% reported having neurological conditions including stroke, transient ischemic attack and Parkinson's disease, 2% reported having major sensory impairments (visual, auditory, and vestibular impairments), and 13% reported having multiple health conditions (i.e., two or more medical conditions). The remaining 21.5% of the sample had cognitive impairments which had been confirmed by clinical examinations (i.e., dementia, age-associated memory impairment, and other memory complaints); 27.5% of these cognitively impaired subjects also reported having other medical conditions (e.g., osteoarthritis, rheumatoid arthritis, musculoskeletal problems, and sensory impairments).

This sample of convenience may not be representative of the elderly community-dwelling population as a whole. However, the proportion of our subjects with and without cognitive impairments (i.e., 21.5% versus 78.5%) closely reflects the 20% prevalence rate of cognitive impairment in community-dwelling elderly that was recently reported by Graham and colleagues (1997). However, our sample may be somewhat over-represented by healthy elderly. MacKinnon (1991) reported that 77% of the men and 80% of the women aged 65 and older who responded to a Canadian General Social Survey in 1985 reported having at least one health problem. In contrast, only 64% of our male subjects and 63% of the female subjects reported having a health problem which interfered with their everyday functioning. These differences, however, may reflect our decision to include subjects aged 60 to 64 who are less likely than their older peers to experience health problems (Chappell, Strain, & Blandford, 1986).

The OARS ADL Scale (Fillenbaum, 1985, 1988) was administered to each of the subjects in their own homes by one of 14 occupational therapists. Most subjects were interviewed directly and provided self-reports. Self-reports of subjects who were diagnosed as having dementia or memory impairment (*n*=58), as well as those subjects with multiple medical conditions which included dementia or memory impairment (*n*=22), were corroborated by a spouse or family member who had regular contact with the subject. Subjects or their informants rated the subject's ability to perform seven PADL items (eating, dressing, grooming, walking, getting in and out of bed, taking a bath or shower, and being able to get to the bathroom on time) and seven IADL items (telephone use, travel, shopping, meal preparation, housework, taking own medications, and handling personal finances). All items, with the exception of the continence item, were rated according to the level of help needed to perform the task (i.e., 2=can perform without help, 1=can perform with some help, and 0=is completely unable to perform the task). The continence item was rated on a different 3-point scale (i.e., 2=no difficulty getting to the bathroom on time or only wets or soils self once or twice a week; 1=problems getting to the bathroom three or more times a week; and 0=requires a catheter or colostomy). A few respondents did not answer all of the questions; this data was treated as missing data. One of the advantages of Rasch measurement approaches is that subjects' data can still be analyzed despite missing data for some subjects (Fisher, Harvey, & Kilgore, 1995; Wright, Linacre, & Heinemann, 1993).

An independent occupational therapy work-up was also completed with each subject at the time of the study. In addition to administering the OARS ADL scale, subjects were interviewed and administered the Assessment of Motor and Process Skills (AMPS) (Fisher, 1997), a semi-individualized observational measure of ADL competence. All available information was then used by the occupational therapist examiners when they completed a summary rating of the subjects' general functional level (i.e., 3=independent; 2=requires minimal assistance to live in the community; 1=requires moderate to maximal assistance to live in the community). Over half of the subjects (54%) were judged by the occupational therapist examiners to require assistance to live in the community (i.e., 29% were expected to require minimal levels of assistance and 25% were expected to require moderate/maximal levels of assistance).

ANALYSES

BIGSTEPS, a Rasch model computer program (Linacre & Wright, 1994b), was used to generate item and person separation statistics, interval-scaled item difficulty measures, person ability measures, and associated goodness-of-fit statistics from the ordinal raw OARS PADL and IADL scores. The hierarchical ordering of the items along the linear continuum of difficulty was examined to determine whether the placement of the items was logical. Unidimensionality of the OARS ADL scale was determined by examining *MnSq* and associated *z* fit statistics for each of the 14 items. Separation statistics were examined to determine whether the OARS ADL scale items separated the subjects into levels of ability (person separation) and whether subjects separated items into levels of difficulty (item separation). Greater separation suggests increased sensitivity of the scale. Subjects' summary functional levels were used to determine if the subjects were logically placed along the linear continuum of ability. The *MnSq* and associated *z* for each subject's calibrated ability measure were examined to determine person response validity. If more than 5% of the subjects' failed to demonstrate goodness-of-fit with the measurement model when *z* is set at 2, then the validity of the OARS ADL scale would be suspect (Wright & Masters, 1982; Wright & Stone, 1979).

Of the original sample of 372 subjects, 51% (*n*=109) received maximum scores on all of the 14 OARS ADL items. These subjects' data did not contribute to the generation of item difficulty calibrations. Consequently, all statistics generated by the BIGSTEPS computer program are based on the data of 182 subjects who scored less than 28 on the OARS ADL scale. Compared to the subjects whose data were analyzed, those subjects who received maximum scores on the OARS ADL items were significantly younger (*M*=72 yrs±7 versus *M*=77 yrs±8)(t=-7.0, p<.001). Maximum scores were more often reported by males (59% of males versus 47% of females) (χ^2=5.1, p=.02), by those with no identifiable medical conditions (80% of those with no identifiable medical condition versus 34% of those reporting at least one medical condition) (χ^2=70.4, p<.001), and by those judged by the occupational therapist examiners as able to live independently in the community (79% of those rated as independent versus 27% of those rated as requiring at least minimal levels of assistance) (χ^2=114.9, p<.001).

RESULTS AND DISCUSSION

Fit of the Items to the Measurement Model

As can be seen in Figure 1, the 14 OARS ADL scale items are spread out over a relatively long continuum of ability from 2.4 to -3.6 logits (see Table 1). The item separation statistic of 6.3 indicates that the OARS ADL items define a long line comprising approximately nine item difficulty strata (i.e., [4 X 6.3 + 1]/3 = 8.7 strata) (Wright & Masters, 1982). At several points on the scale, two items are located within 0.25 logits of one another (i.e., "money management" and "meal preparation," "transportation" and "medication management," and "walking" and "grooming").

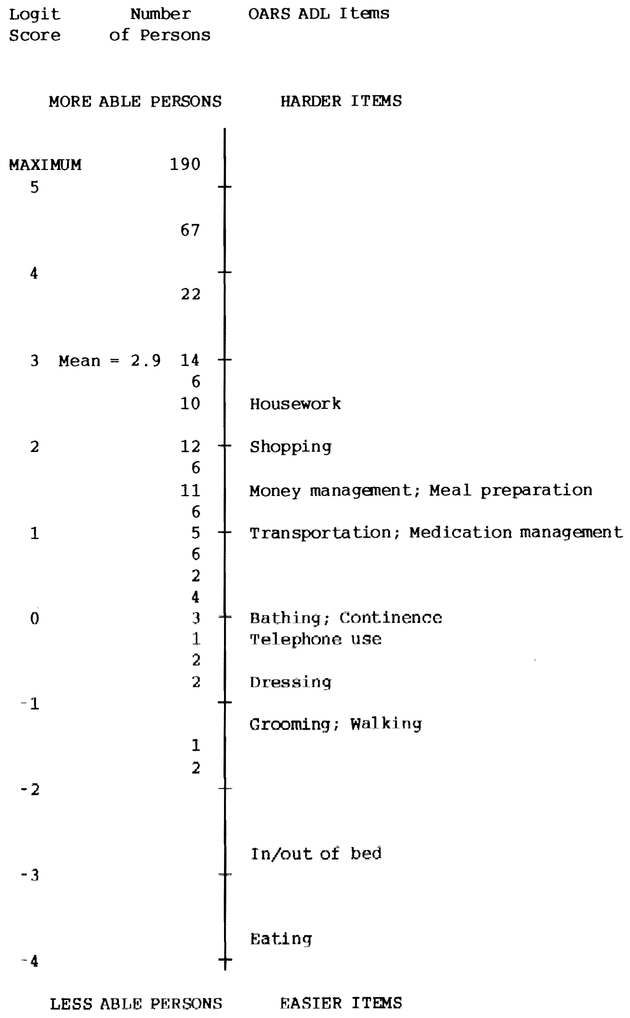

FIGURE 1 Location of persons and items on the linear continuum of ability/difficulty.

Table 1
Item Difficulty Calibrations and Fit Statistics for 14 OARS ADL Items (*n* = 182)

ADL/IALD Items	Item difficulty Measure	Error	Infit MnSq	*z*	Outfit MnSq	*z*
Housework	2.36	.14	1.0	-1	1.0	0
Shopping	1.94	.14	0.6	-4	0.6	-4
Money management	1.50	.15	1.0	1	0.9	-1
Meal preparation	1.30	.15	0.9	-1	0.8	-2
Transportation	1.09	.16	0.7	-3	0.7	-2
Taking medication	0.84	.16	0.9	-1	0.8	-1
Bathing	0.45	.17	1.3	3	1.3	1
Continence	0.12	.19	2.0*	6	4.5*	7
Using phone	-0.16	.20	0.8	-2	0.5	-2
Dressing	-0.72	.22	0.7	-2	0.4	-2
Grooming	-1.18	.25	0.9	0	0.8	0
Walking	-1.24	.26	1.1	0	0.7	-1
In/out of bed	-2.66	.41	1.0	0	3.0	1
Eating	-3.63	.60	0.9	0	0.3	-1

* MnSq residuals >1.4 with standardized *z* scores ≥ 2 signify a failure to demonstrate goodness-of-fit with the Rasch measurement model.

The 14 items are positioned along the hierarchical continuum in a logical order. The item order is generally consistent with the findings of other research studies (Finch et al., 1995; Fillenbaum, 1988; Kempen & Suurmeijer, 1990; Siu et al., 1990; Silverstein et al., 1992; Spector et al., 1987; Suurmeijer et al., 1994) and with Lawton and Brody's (1969) theoretical conceptualization of the relative difficulty of PADL and IADL items. However, given that some items are of similar difficulty (e.g., money management and meal preparation, and transportation and medication management), the actual order of the items may vary when other samples are assessed. IADL items such as housework, shopping, money management, and meal preparation are more difficult than PADL items such as eating, getting in and out of bed, walking, and grooming. There is, however, some evidence of overlap between the two types of items. Two PADL items, bathing and continence, are more difficult than the IADL item of using the telephone. However, large gaps among the 9 strata of items are evident. Of particular concern is the paucity of items calibrated as more difficult than 2.0 logits. Fifty-one percent of the original sample of 372 subjects received maximum scores, even on the most difficult items (i.e., housework and shopping). The development of more difficult items such as washing windows, mowing the lawn, putting out the garbage, vacuuming carpets, and mopping the floor may serve to improve the sensitivity of the OARS ADL scale.It may also be useful to create several items from broadly defined items such as money management and meal preparation. For example, meal preparation could be conceptualized as consisting of a relatively easy item such as preparing a snack or heating something up in the oven or microwave oven, as well as a more difficult item such as preparing a hot meal using a stove or oven. A longer scale comprising more closely spaced items would enhance our ability to identify those older persons who are beginning to show early signs of functional decline. Even if early signs of functional decline do not indicate an immediate need for external supports and services, these signs may enable health professionals to identify persons who may be at risk in the near future.

Gaps along the lower third of the scale (i.e., below -1.0 logits) also indicate a lack of easier items although this may be of little practical concern. The very large estimated errors for the two easiest items eating and getting in and out of bed (i.e., .60 and .41 respectively) (see Table 1) indicate that these items were too easy for our sample. Even the most unable subject in our sample was able to perform these two easy items independently. By the time community-dwelling elderly require the assistance of another person to eat and to get in and out of bed, they are probably already receiving full-time supervision within their home or in a long-term care facility.

The associated goodness-of-fit statistics for each of the 14 OARS ADL item difficulty calibrations are listed in Table 1. Thirteen of the 14 items demonstrated acceptable goodness-of-fit with the measurement model (infit and outfit *MnSq* values ≤ 1.4 with $z<2$). The "continence" item demonstrated unacceptable infit and outfit *MnSq* values (infit *MnSq*=2.0, z=6, outfit *MnSq*=4.5, z=7). Furthermore, 18% (n=33) of the subjects' responses to the continence item were unexpected. Of the 28 subjects who scored lower than expected, 15 reported being healthy and seven reported having a medical condition such as osteoarthritis which would not be expected to adversely affect their bladder functioning. Only five subjects had medical conditions (e.g., diabetes, kidney disease, tumors on the bladder) which might provide an explanation for their continence problems. All five of the subjects who scored higher than expected had intact bowel and bladder functioning even though they functioned poorly on many other ADL items as a result of Alzheimer's disease (n=4), or visual impairments and back problems (n=1). Linacre and colleagues (1994) similarly found that the bladder management item of the Functional Independent Measure (Granger, Hamilton, & Sherwin, 1986) failed to demonstrate goodness-of-fit with the Rasch measurement model. The inability to manage one's bladder or remain continent is more likely a reflection of the integrity of one's underlying physiological functioning rather than an indication of one's ADL ability (Fisher et al., 1994; Linacre, Heinemann, Wright, Granger, & Hamilton, 1994). The fact that the scoring system used to rate the continence item on the OARS ADL scale is different from that used for all of the other items suggests that the developers of the scale intuitively knew this item was measuring a different construct. New items which measure individuals' ability to adapt to a continence problem (e.g., being able to clean themselves, and launder their clothing and/or bedding) may better represent the underlying construct of ADL ability.

Given the obvious failure of the item "continence" to demonstrate acceptable goodness-of-fit statistics, this item was removed and the data for the remaining 158 subjects who showed some variation on their scores for the other 13 items were reanalyzed. During the second analysis, 12 of the 13 items (i.e., five PADL and seven IADL items) continued to demonstrate acceptable goodness-of-fit statistics. "Bathing", however, failed to demonstrate acceptable goodness-of-fit statistics (infit *MnSq*=1.5, z=3, outfit *MnSq*=1.5, z=2). Furthermore, 9.5% (n=15) of the sample scored unexpectedly on this item. All 15 subjects rated their ability to bathe lower than expected. Most of these subjects (n=13) reported having medical conditions that affected their strength, balance, and joint mobility, particularly the mobility of their hip joints. These problems may affect the confidence they have in their ability to get in and out of the bathtub without falling or sustaining an injury. Two subjects who reported being physically healthy, obtained assistance from family members or community nurses to reduce the risk of falling when getting in and out of the bathtub. As currently worded, the bathing item appears to measure subjects' ability and confidence in getting in and out of the bathtub rather than other aspects of bathing such as their ability to wash and dry themselves. Constructing items which address these issues separately may enable us to determine if one or both of the revised items demonstrate more acceptable fit with the measurement model.

During a subsequent third analysis using the data of 155 subjects, both "continence" and "bathing" were omitted. The item difficulty calibrations for the remaining 12 items are presented in Table 2. All 12 remaining items were positioned on the linear continuum in the same order as during the first and second analyses, and continued to demonstrate acceptable goodness-of-fit statistics. Since these 12 items represent a unidimensional construct of ADL ability, researchers can combine these 5 PADL and 7 IADL items to generate a single measure of ADL ability. The item separation statistic for the 12 item OARS ADL scale was greater than for the 14 item scale (7.3 versus 6.7) and 10.1 distinct strata of item difficulty were identified. Although the length of the line was increased, the sensitivity of the OARS ADL scale was unchanged.

Fit of the Persons to the Measurement Model

The findings support the claim of previous research which has suggested that self-report measures, such as the OARS ADL scale, are limited in their ability to distinguish among persons, especially those at higher levels of functioning (Guralnik, Branch, Cummings, & Curb, 1989). The 14 item OARS ADL scale person separation index of only 0.9 indicates that the OARS ADL scale is able to separate this heterogeneous sample of community-dwelling elderly adults into only two distinct strata of ability levels. The removal of the continence and bathing items had no meaningful effect on the person separation index (i.e., it increased from 0.9 to only 1.1). Generally, the ordering of the subjects along the linear continuum of

Table 2
Item Difficulty Calibrations and Fit Statistics for 12 OARS ADL Items (n = 155)

ADL/IALD Items	Item difficulty Measure	Error	Infit MnSq	*z*	Outfit MnSq	*z*
Housework	2.88	.15	1.1	1	1.2	1
Shopping	2.36	.16	0.8	-2	0.7	-2
Money management	1.82	.16	1.2	1	1.0	0
Meal preparation	1.57	.17	1.0	0	0.9	-1
Transportation	1.31	.17	0.9	-1	0.9	-1
Taking medication	1.01	.18	1.0	0	1.0	0
Using phone	-0.19	.21	0.9	-1	0.6	-1
Dressing	-0.85	.24	0.7	-2	0.5	-1
Grooming	-1.37	.27	1.0	0	1.3	1
Walking	-1.45	.28	1.3	2	1.3	0
In/out of bed	-3.03	.43	1.2	1	6.1	1
Eating	-4.08	.62	0.9	0	0.3	0

ADL ability was reasonable. Subjects assumed to be more able (i.e., those with no identifiable medical conditions) were typically located at the more able end of the linear continuum. Less able subjects (i.e., those with severe impairments including dementia) were located at the other end of the linear continuum. In addition, those who were judged by the occupational therapist examiners to be able to live in the community independently were generally located at the more able end of the linear continuum; those who were judged to require moderate to maximal levels of assistance were usually located at the less able end of the continuum. When placement of subjects on the linear continuum was unexpected, most often subjects were identified as being more able than other available information suggested (i.e., reported medical conditions, AMPS ability measures, and functional levels). For example, 27% of those who achieved maximum scores on the OARS ADL items were judged by the occupational therapist examiners to require at least minimal levels of assistance to live in the community. These subjects may have overestimated their ADL ability.

Wright and Stone (1979) described "a best test" as "one which measures best in the region within which measurements are expected to occur" (p. 133). A well-targeted test should consist of items which are a bit too hard, a bit too easy, and just right for those for whom the test was designed (Wright & Stone, 1979). Consequently, we would expect that the mean ability measure of a target sample would be located around the mean item difficulty calibration (i.e., around 0.0 logits). However, the subjects' mean ability measure of 2.9 logits was well above the mean item difficulty calibration. Consequently, the sample distribution was skewed with 98% of the sample scoring above the mean item difficulty calibration. Clearly, the OARS ADL scale is very poorly targeted to this heterogeneous sample of community-dwelling elderly.

It is apparent from the high proportion of subjects who rated themselves as independent on all 14 items that the items comprising the OARS ADL scale are insufficiently challenging for most community-dwelling elderly. Furthermore, the failure of the OARS ADL scale to adequately differentiate the higher functioning subjects into more discrete ability levels may, in part be attributed to the use of a 3-point rating scale. Changes in the effectiveness and efficiency with which a person performs a task will not be identified when subjects rate their abilities using the current 3-point rating scale. Even if individuals perform a task more slowly, are less organized in their approach to the task, or experience some difficulty overcoming problems that arise during the course of task performances, if they do not require the physical assistance of another person, they will rate their performance as "independent". To enhance the utility of the OARS ADL scale in identifying older persons who while independent, perform less effectively and efficiently, a 4-point scale may be more appropriate (e.g., 3=can perform the task with ease; 2=experiences some difficulty but can still perform the task independently; 1=requires the help of another person to perform the task; 0=another person must perform the task for the individual).

The majority of the subjects demonstrated acceptable fit with the measurement model whether 14 or 12 items were examined. When all 14 items of the OARS ADL scale were analyzed, subject fit was 97%. Subject fit increased to 99% when continence and bathing were omitted. The one subject whose responses failed to demonstrate acceptable goodness-of-fit was located at the less able end of the linear continuum of ADL ability and scored lower than expected on the easiest item (i.e., "eating").

CONCLUSIONS

The use of Rasch measurement enables us to transform ordinal OARS ADL ratings for each subject to an interval measure of person ability. The utility of an ADL ability measure is, however, dependent on the items that comprise the test. We found that two of the OARS ADL items, namely continence and bathing, did not demonstrate acceptable goodness-of-fit with the measurement model. The remaining five PADL and seven IADL items, however, are hierarchically ordered, measure the same underlying construct (i.e., perceived ADL ability), and thus can be combined to generate a single ADL ability measure for subjects. This finding argues against the usual practice of generating separate PADL and IADL scores for subjects. The use of ability measures based on the 12 items that demonstrate goodness-of-fit with the measurement model will enable us to make more accurate comparisons of individuals' ADL ability over time, and between different individuals. The validity of the ADL ability measures, however, is markedly limited by the finding that the OARS ADL items are poorly targeted to the community-dwelling elderly for whom the OARS was designed. The large gaps between items, and the lack of items at both ends of the linear continuum, especially at the more difficult end where the majority of our sample was located, and failure of the assessment to meaningfully separate the sample on the basis of ADL ability, prevent us from being able to adequately differentiate subjects on the basis of ADL ability.

We have suggested several possible revisions to increase the length of the scale, reduce the number of gaps between items, and thereby improve the overall targeting and sensitivity of the test to changes in subjects' ADL abilities. Only then will the ability measures generated by the OARS ADL scale be appropriate for use in outcome and longitudinal studies.

ACKNOWLEDGMENTS

This study was supported by grants from the Alzheimer's Association, the Alzheimer Society of Canada, and the National Institute on Aging, National Institutes of Health, United States Department of Health and Human Services. Financial support for the doctoral studies of the first author was received through the Thelma Cardwell Scholarship from the Canadian Occupational Therapy Foundation, a Gerontology Scholarship from the Royal Canadian Legion, and doctoral scholarships from the Medical Research Council of Canada, National Health and Research Development Programme, and the Alzheimer Society of Canada. Appreciation is extended to Norma Lewis who assisted in preparing the data for analysis and to Dr. John Fisk who provided valuable feedback and editorial assistance.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders DSM-IV*. Washington, DC: American Psychiatric Association.

Branch, L. G. & Ku, L. (1989). Transition probabilities to dependency, institutionalization, and death among elderly over a decade. *Journal of Aging and Health, 1*, 370-408.

Chappell, N., Strain, L., & Blandford, A. (1986). *Aging and health care: A social perspective*. Toronto: Holt, Rinehart and Winston of Canada.

Fillenbaum, G. G. (1988). *Multidimensional functional assessment of older adults: The Duke Older Americans Resources and Services Procedures*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Finch, M., Kane, R. L., & Philp, I. (1994). Developing a new metric for ADLs. *Journal of the American Geriatrics Society, 43*, 877-884.

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

Fisher, A. G., Bryze, K. A., Granger, C. V., Haley, S. M., Hamilton, B. B., Heinemann, A. W., Puderbaugh, J., Linacre, J. M., Ludlow, L. H., McCabe, M.

A., & Wright, B. D. (1994). Applications of conjoint measurement to the development of functional assessments. *International Journal of Educational Research, 21*, 579-593.

Fisher, W. P. (1993). Measurement-related problems in functional assessment. *American Journal of Occupational Therapy, 47*, 331-338.

Fisher, W. P., & Fisher, A. G. (1993). Applications of Rasch analysis to studies in occupational therapy. *Physical Medicine and Rehabilitation Clinics of North America, 4*, 551-569.

Fisher, W. P., Harvey, R. F., & Kilgore, K. M. (1995). New developments in functional assessment: Probabilistic models for gold standards. *NeuroRehabilitation, 5*, 3-25.

Gillen, P., Spore, D., More, V., & Freiberger, W. (1996). Functional and residential status transitions among nursing home residents. *Journal of Gerontology, 51A*, M29-M36.

Graham, J., Rockwood, K., Beattie, B. L., Eastwood, R., Gauthier, S., Tuokko, H., & McDowell, I. (1997). Prevalence and severity of cognitive impairment with and without dementia in an elderly population. *Lancet*, 1793-1796.

Granger, C. V., Hamilton, B. B., & Sherwin, F. S. (1986). *Guide for Use of the Uniform Data Set for Medical Rehabilitation*. Buffalo: Buffalo General Hospital.

Greiner, P. A., Snowdon, D. A., & Greiner, L. H. (1996). The relationship of self-rated function and self-rated health to concurrent functional ability, functional decline, and mortality: Findings from the Nun study. *Journal of Gerontology, 51B*, S234-S241.

Guralnik, J. M., Branch, L. G., Cummings, S. R., & Curb, D. J. (1989). Physical performance measures in aging research. *Journal of Gerontology, 44*, M141-M146.

Guttman, L. (1950). The basis for scalogram analysis. In S.A. Stouffer et al. (Eds), *Studies in social psychology in World War II* (Vol. 4, Measurement and prediction). New York: Wiley.

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.

Katz, S., Ford, A. B., Moskowitz, R. W., Jackson, B. A., & Jaffe, M. W. (1963). Studies of illness in the aged. The Index of ADL: A standardized measure of biological and psychosocial function. *Journal of the American Medical Association, 185*, 914-919.

Kempen, G. I. J. M. & Suurmeijer, T. P. B. M. (1990). The development of a hierarchical polychotomous ADL-IADL scale for noninstitutionalized elders. *Gerontologist, 30*, 497-502.

Lawton, M. P. & Brody, E. M. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. *Gerontologist, 9*, 179-186.

Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V., & Hamilton, B. B.

(1994). The structure and stability of the functional independence measure. *Archives of Physical Medicine and Rehabilitation*, *75*, 127-132.

Linacre, J. M., & Wright, B. D. (1994a). Chi-square fit statistics. *Rasch Measurement Transactions*, *8*, 360-361.

Linacre, J. M., & Wright, B. D. (1994b). *A user's guide to BIGSTEPS: Rasch-model computer program (version 2.4)*. Chicago: MESA Press.

MacKinnon, A. L. (1991). Occupational performance of activities of daily living among elderly Canadians in the community. *Canadian Journal of Occupational Therapy*, *58*, 60-66.

Manton, K. G. (1988). A longitudinal study of functional change and mortality in the United States. *Journal of Gerontology*, *43*, S153-S161.

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation*, *70*, 308-332.

Pfeiffer, B. A., McClelland, T., & Lawson, J. (1989). Use of the functional assessment inventory to distinguish among the rural elderly in five service settings. *Journal of the American Geriatrics Society*, *37*, 243-248.

Poon, L. W. (1994). On the paradox of improving sensitivity of ADL scales for the detection of behavioral changes in early dementia. *International Psychogeriatrics*, *6*, 171-177.

Rubenstein, L. Z., Schairer, C., Wieland, G. D., & Kane, R. (1984). Systematic biases in functional status assessment of elderly adults: Effects of different data sources. *Journal of Gerontology*, *39*, 686-691.

Silverstein, B. S., Fisher, W. P., Kilgore, K. M., Harley, J. P., & Harvey, R. F. (1992). Applying psychometric criteria to functional assessment in medical rehabilitation: II. Defining interval measures. *Archives of Physical Medicine and Rehabilitation*, *73*, 507-518.

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

Spector, W. D., Katz, S., Murphy, J. B., & Fulton, J. P. (1987). The hierarchical relationship between activities of daily living and instrumental activities of daily living. *Journal of Chronic Diseases*, *40*, 481-489.

Suurmeijer, T. P. B. M., Douglas, D. M., Mown, T., Briançon, S., Krol, B., Sanderman, R., Guillemin, F., Bjelle, A., & van den Heuvel, W. J. A. (1994). The Groningen activity restriction scale for measuring disability: Its utility in international comparisons. *American Journal of Public Health*, *84*, 1270-1273.

Wilson, M. (1989). A comparison of deterministic and probabilistic approaches to measuring learning structures. *Australian Journal of Education*, *33*, 127-144.

Wolinsky, F. D., Callahan, C. M., Fitzgerald, J. F., & Johnson, R. F. (1993). Changes in functional status and the risks of subsequent nursing home placement and death. *Journal of Gerontology*, *48*, S94-S101.

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

Wright, B. D., Linacre, J. M., & Heinemann, A. W. (1993). Measuring functional status in rehabilitation. *Physical Medicine and Rehabilitation Clinics of North America*, *4*, 475-491.

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

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

JOURNAL OF OUTCOME MEASUREMENT®, *2*(1), 25 - 32

Man is the measure ... the measurer

Mark H. Stone
Adler Institute

Measures originated from human anatomy. Metrology has moved from man the measure to man the measurer. This transformation is documented using examples taken from the history of metrology. The outcome measures are units constructed and maintained for their utility, constancy and generality.

Requests for reprints should be sent to Mark H. Stone, Adler School of Professional Psychology, 65 East Wacker Place, Suite 2100, Chicago, IL 60601-7203.

Socrates, early in the Theaetetus (Cornford, 1957), opens the discussion on knowledge saying, "He [Proagoras] says, you will remember, that 'man is the measure of all things.' " (p.31). Protagoras originally made his famous statement 100 years before Plato who, using the words of Socrates, has perpetuated it for posterity. Man is the measure. So begins the process of measurement. From man come the units. Our earliest knowedge of anything is grounded in reality and it is the attributes of the human body that first give us useful units, the [finger] digit and the foot. However, Socrates, later in the dialogue, goes on to point out that making man the measure leads to a proliferation of individual perceptions, a state which cannot produce exact knowledge. Using man as the measure produces a variation to each human unit. Therefore, much more work is required to achieve measures that will be useful. I will return to this point later.

MAN THE MEASURE

The beginnings of metrology are anthropomorphic. The first units came from the human body. Naturally prominent among these anatomic units is the finger digit and the foot, the origins of today's inch and foot.

The Inch

Inch derives from "onyx" or "onych" which is Greek for nail or claw. Onyxis is the medical term for an ingrown nail. The dimensions of the digit "inch" were usually taken across the finger or thumb at the level of the nail. But how was the variance of digit "inches" addressed?

"The thowmys of iii men, that is to say a mekill man, and a man of messurabel statur, and of a lytell man. The thoumys are to be mesouret at the rut of the nayll" (Nicholson, 1912, p. 35). Written at the time of David I of Scotland c. 1150, this quote indicates how a "standard" inch was achieved. First, there is a standard method by which the measure is made i.e. the measure is taken across the digit at the nail. Second, variation is recognized in such measures even though care is given as to how the measure is made. Variation is addressed by "averaging" the results of a large (mekill), middle (messurabel) and little (lytell) man. This quote indicates a solution to finding the "average" width of a human attribute. It can be obtained by considering the extremes values (mekill and lytell) in addition to a middle value (messurabel). This gives a useful solution to the problem of what values to use in determining a unit length. Third, *man* is the measure. While initially appearing as sexist bias, it actually serves a useful purpose because restricting the sample to only one sex serves to further reduce variation. Consequently, we find in this early reference a recognition of the problem of establishing a standard "inch." The variation that comes from considering man the measure is resolved by following a simple sampling plan that produces a solution to the problem.

The Foot

An etching made in the 16th century shows 16 men standing in a line outside the entrance to a church. Each man has placed his left foot heel-to-toe with another. Observing the scene are three monitors. The accompanying narrative to the etching describes the method being employed. The narration indicates that monitors were stationed at the door of the church following the weekly service to

> bid sixteen men to stop, tall ones and short ones, as they happened to come out ... their left feet one behind the other ... gives ... the right and lowful rood ... [the sixteenth part of which is] ... the right and lawful foot. (Nicholson, 1912, p. 47)

In a manner similar to the determination of the inch, sixteen men exiting from church are commandeered to provide a sample from which "the sixteenth part of which" is used to established a standard foot.

While it is customary to view early historical accounts as quaint and primitive compared to modern metrology, such a viewpoint fails to note the ingenuity inherent in these simple solutions. These early determinations of the unit inch and foot are quite sensible. These two examples reveal a thoughtful solution for determining units derived from the human body. The variation inherent in human physiology is recognized and a unit is produced by following a method to determine an average that can be adopted as a standard. Early metrology recognized variation and devised ways to address it.

The rod (16 feet), commonly used in surveying, is made from the total sample of men and their 16 successive left feet. From one strategy, two useful units are derived.

MAN THE MEASURER

From Das Feurwerksbuch, c. 1420 we find this admonition, "A Master ... first of all should know three things — the weights of solids, the weights of liquids, and the methods of measuring" (Nicholson, 1912, p. 67). The methods of measuring, man in the two examples cited, demonstrate careful attention to the measuring process.

> Leonardo daVinci wrote, . . . Vittruvius [Pollio, 1st century] declares that nature has thus arranged the measurements of a man: four fingers make one palm, and four palms make one foot; six palms make one cubit; four cubits make once a man's height; four cubits make a pace; and twenty-four palms make a man's height. (Berrimman, 1953, p. 87)

For a 6-foot man we have these measures:

Unit	Inches	Method
finger	0.75	at the nail,
palm	3	across 4 fingers excluding thumb,
foot	12	from among 16 men,
cubit	18	forearm, elbow to middle finger,
span	66	arm to arm across the chest,
man's height	72	in an upright stance,
man's pace	72	left-right-left cycle.

This system of units did not arise haphazardly. It was the consequence of much attention to the problems inherent in making measures. The craft of measurement was given much attention and this arrangement of anthropomorphic units testifies to it.

Mention of the cubit (elbow to the middle finger) deserves some comment. This unit has been used from biblical times to the present and, even now, is still used in parts of Asia. Macdonell (1902) used the cubit as one of seven body measurements in a study of criminals. He and Pearson studied this data in the Galton Laboratory precipitating work on analysis of the principal axes of the correlation ellipsoid.

The cubit was commonly used in all the countries around the Mediterranean during the Greek and Roman empires. The cubit is about half of our yard and occupies a position in the unit scheme that has not been filled in modern times, although sometimes you will encounter an 18-inch ruler. Eighteen-inch rulers are, incidentally, handy measuring tools, especially when you have to draw a 12-inch line! The common, 12-inch ruler is too short for measuring and drawing lines in the range of one-foot. Because the 12-inch ruler fits more conveniently into cases and bags, it may be one of the reasons why a 12-inch ruler continues to be used rather than an 18-inch one. Due to the invention of measuring tape, the cubit will probably never return to its former glory. Nevertheless, the cubit remains with us today. It goes wherever we go!

The range of lengths for the cubit is from 17 inches to 25 inches with most of the values between 18 and 20 inches. My cubit is 20 inches. What is yours? How shall we resolve the difference. Compute an average the way the monitors did with the unit foot.

Mention of cubits brings us to biblical references, and you will find the cubit mentioned in Exodus 25:10, I Kings 6:25ff. and Nehemiah 3:13 where it specifies linear measures of buildings and objects.

References in the Bible concerning measurement focus upon two other important measurement devices — scales and weights. The scale or balance beam is, perhaps, the oldest method known for weighing objects. It was indispensable to early commerce and still provides the device needed to facilitate fair trade and just recompense. Justice is an important biblical issue and the prophets spoke of its preeminence. The "scales of justice" remain to this day as a contemporary icon.

In Leviticus 19:35 it is written, "You shall do no wrong in judgments in measures of length or weight or quantity" (RSV). Micah 6:11 adds, "Shall I acquit the man with wicked scales and with a bag of deceitful weights?" (RSV).

To assure justice, the balance beam can be monitored by reversing the arm and also by requiring that the contents of each side be exchanged. But what of the weights themselves? This problem is as old as man because, in fact, the problem is man! The deceitful strategy, complained of by the prophets, was to use two sets of weights, one set hidden from view. The heavier set was used for purchasing and the lighter set for selling. The trick was to palm one set out of view and substitute the other depending upon whether one was buying or selling. Small wonder that the prophets raged about the evil contained in this practice. Even the wrath of God was implored to assure just measures. In Deuteronomy 25:13-16 we find written,

> You shall not have in your bag two kinds of weights, a large and a small. You shall not have in your house two kinds of measures, a large and a small. A full and just weight you shall have, a full and just measure you shall have; that your days may be prolonged in the land which the Lord your God gives you.For all who do such things, all who act dishonestly, are an abomination to the Lord your God. (RSV)

Unfortunately, the evil intentions of man were not curbed by invoking the wrath of God. Deceit in measuring continues.

One measuring abuse was perpetuated through the divine right of kings, especially for procuring more taxes. In the 35th of the 63 clauses contained in the Magna Charta signed by King John on 15 June 1215 at Runnimede we find the following,

> Throughout the kingdom there shall be standard measures of wine, ale, and corn. Also there shall be a standard width of dyed cloth, russet, and haberject; namely two ells with the selvedges. Weights are to standardized similarly. (Berriman, 1953, p. 560)

Russet and habeject are homespun cloth. Haberject is derived from the same root word from which we get haberdasher. An ell is about 45 inches and selvedges is the word for edges. Most important, we find, in this agreement negotiated with the king, specifications about standard units. We take for granted the absolute necessity of standard units without which commerce would be impossible. Today, as in the past, concern for units arises when we feel cheated.

Recite the "Jack and Jill" of your youth and give close attention to what you say. "Jack" (King John) fell, the "jill" tumbled and the "crown" was broken. Now we understand the Magna Charta anew. As in most nursery rhymes and fairy tales, there is more contained in the story than what is first heard in its recitation.

Life was hard in those days and the units reflect this. The proverbial "hand to mouth existence" is nowhere more evident than in the units of past times — mouthful, handful (a jigger and twice the mouthful), jack (from which we get jackpot, and double jigger), gill (or jill which is a double jack), cup (a double jill) and pint (a double cup). Observe the part that alcohol plays in the metrology of man, particularly for taxation.

There are many illustrations that could be considered in this story of units taken from human anatomy, but we have enough in these examples to draw some conclusions to the story of man the measure, and man the measurer.

1. Measures began with units derived from experience. Human anatomy provided the first units. How much closer to experience can you get? The hand, arm, foot and mouthful are units common to every person. These units are empowered by their utility and generality.

2. Fairness and deceit have always been a part of measuring. Early times were as plagued by this problem as we are today. Progress in measuring has required constant vigilance against willful intrusion and ma-

nipulation in the making of measures. Just measuring requires the elimination of all such intrusions or, at the very least, controling their influence.

3. Using man to derive units appears sexist, but it reduced variation. Neither gender should play a part in making measures, but using units from only one sex kept the values from even greater variation. Single-sex units served a purpose.

4. Progress has evolved from man as the measure, i.e. a unit of human anatomy, to man as the measurer. As Oliver Goldsmith wrote in The Goodnatured Man (1768), "Measures, not men have always been my mark."

5. We have progressed from units taken from anatomy to "absolute" units constructed and maintained for their utility, constancy and generality. We have moved from fact to fiction, direct to indirect, concrete to abstract and manifest to latent.

OUTCOME MEASUREMENT

How does this story of early metrology relate to outcome measurement? Fundamental in making measures is the construction of as variable and determination of an origin and a unit. A variable is constructed from ordered items shown to have imputed and verified relevance. The origin and unit employed are arbitrary. The origin is usually taken from some standard such as a column of water at sea level or else it is set at zero. The unit is typically derived from experience, but later it is usually abstracted from reality as in the examples of the inch and foot.

Unit monitoring is indispensable to achieving useful measures. All measuring tools require consistency, but consistency can only come through continuous monitoring. We make measures successfully only when such conditions have been met. Outcome measurement rests upon the fundamental considerations observed in these examples taken from the history of metrology which illustrate the transformation from man as the measure to man the measurer, from experience to method, from reality to an abstraction. This abstraction is essential in building scales for outcome measurement. As Russell Fox (1963) wrote, "Measurement has meaning only if we can transmit the information without ambiguity to others" (p. 163).

REFERENCES

Berriman, A. E. (1953). *Historical metrology.* London: Dent & Sons.

Barrell, H. Metrology. In *The new encyclopedia Britannica (Vol. 15)*, pp. 310-314. Chicago: Encyclopedia Britannica.

Cornford, F. M. (1957). *Theaetetus: Plato's theory of knowledge.* New York: Macmillan.

Dresner, S. (1971). Units of measurement. London: Aylesbury and Medcalf.

Fox, R. (1963). *The science of science.* New York: Westinghouse.

Fowler, R, & Meyer, D. (1961). *Physics for engineers and scientists.* Boston: Allyn and Bacon.

Lindsell, H. (Ed.). (1964). *Harper Study Bible, Revised Standard Version.* Grand Rapids, MI: Zondervan.

Macdonell, W. (1902). On criminal anthropometry and the identification of criminals. *Biometrika, 1*, 177-227.

Nicholson, E. (1912). *Men and measures.* London: Smith & Elder.

JOURNAL OF OUTCOME MEASUREMENT®, *2*(1), 33 - 42

Evidence for the Validity of a Rasch Model Technique for Identifying Differential Item Functioning

Janice Dowd Scheuneman
Raja G. Subhiyah
National Board of Medical Examiners

This paper presents an analysis of differential item functioning (DIF) in a certification examination for a medical specialty. The groups analyzed were (1) physicians from different subspecialties within this area and (2) physicians who qualified for the examination through two different experiential pathways. The DIF analyses were performed using a simple Rasch model procedure. The results were shown to be readily interpretable in terms of the known differences between the groups being compared. These results serve as validity evidence for the Rasch model procedure as a means for evaluating DIF in examinations. The conclusion is drawn that complex procedures are not required to generate interpretable results if relevant differences between the groups being compared are known. This suggests that the inability of many researchers to interpret results for racial/ethnic or gender groups is not due to inadequacies of the methods, but more likely to lack of pertinent knowledge about group differences.

Requests for reprints should be sent to Janice Dowd Scheuneman, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104.

More than 20 years ago, the issues of what was then called "item bias" first began to be discussed. These issues arose out of concern for the observed differences between major demographic groups, such as those defined by race or ethnicity, in their performance on a wide variety of educational examinations. At that time, many researchers assumed that once a method for identifying biased items was established, they could quickly move on to determine what caused this phenomenon, develop guidelines for item writers to avoid these causes, and revise tests accordingly. Unfortunately, as more and more studies of item bias were performed, the primary result for most researchers was that they were unable to identify features of items that might account for the statistical findings. Moreover, items identified by expert reviewers as presenting potential difficulties for the groups being studied failed to correspond well with those identified by the item bias methods.

Two explanations might be offered for these seemingly anomalous results. The first explanation is that the methods used in these studies may not be identifying the right items. That is, if the methods were sufficiently accurate, the results would be interpretable. The second explanation is that most of the proposed methods do perform adequately to identify a pool of items that might be examined for possible sources of item bias or differential item functioning (DIF), as it is now more commonly termed. The inability to interpret results stems instead from sources such as (a) inadequate theory of the cognitive interaction between person and test item; (b) inadequate knowledge of the interaction of culture and characteristic modes of learning and comprehension; and (c) inadequate models for properties of items that are associated with their difficulty or for how differential item difficulty might arise.

As the difficulties in interpretation have become increasingly apparent, the majority of researchers have chosen to pursue the first of these alternatives. Many procedures of increasing elegance and refinement have been developed and reported in the measurement literature over the intervening years (See Berk, 1982; Hills, 1989; Rudner, Getson, & Knight, 1980a; Scheuneman & Bleistein, 1989; Shealy & Stout, 1993; Swaminathan & Rogers, 1990). Unfortunately this technical progress has not led to corresponding progress in our understanding of the mechanisms that produce DIF. With the exception of a few isolated examples, the reasons for the results remain murky for the majority of tests that have been studied.

Our hypothesis is that if the differences between the groups to be analyzed are well understood, even a simple procedure is adequate to produce interpretable and useful results. For this purpose, we selected a certifying examination for a medical specialty area for which examinee groups may be clearly differentiated according to their primary areas of expertise within the specialty or according to the pathway for qualifying to sit for the examination. The method chosen is one of the earliest simple procedures, the Rasch model method suggested by Draba (1977). This method did not perform well in early studies comparing the techniques then being proposed (e.g., Rudner, Getson, & Knight, 1980b; Shepard, Camilli & Averill, 1981) and has received little attention since that time although the conditions under which it was evaluated in those studies were not those where Rasch procedures might have been expected to work well. Particularly for tests where the Rasch model is used for other purposes, as it was for the examination studied, this procedure seems appropriate and easy to use. Moreover, these procedures are appropriate for use with sample sizes smaller than are generally recommended for use of the more sophisticated methods.

METHOD

Data Source

The data were taken from the certifying examination for medical specialty that consisted of two parts: a 250-item core examination that must be taken by all candidates and three 125-item specialty area examinations, one of which must be selected by each candidate as most appropriate for his/her training. Candidates qualify for the examination either by successfully completing a recognized residency program or by demonstrating the acquisition of equivalent experience in their practice. The study was based on the responses of 416 candidates to 244 items on the core examination. (Six of the original 250 items were deleted from scoring following administration of the examination in 1993.)

Two important examinee characteristics were used to identify sets of comparison groups. The first set of groups was identified by qualification pathway: residency vs. equivalency. The second was identified by specialty area examination selected by the candidates. One of these specialty areas had only 41 candidates, however, so that specialty area analyses were performed only for the two other groups. The candidates from the third specialty area were included in the analyses by pathway, however.

Procedure

Separate Rasch item-calibration analyses using BIGSTEPS software (Wright & Linacre, 1992) were done on the 244 items on the core examination for each of the four groups of candidates and for two randomly selected baseline groups. Rasch difficulty estimates were compared item by item across groups in the two sets of interest and across the random groups. Based on the recommendation by Draba (1977), an item was identified as exhibiting DIF if the difference between an item's difficulty estimates for the two groups was more than .50 logits.

Difficulty estimates were then compared across the two random groups to get a baseline frequency with which DIF was identified and to confirm the suggested cut-off value of .50 for difficulty differences. Only seven items (2.9%) were identified when these analyses were performed. Since this result was satisfactory, it was decided to accept the value .50 as criterion for the other comparisons. Rasch item difficulty estimates for the residency-equivalency and the specialty area groups were then compared.

The same comparison groups were subsequently analyzed using a Mantel-Haenszel (M-H) procedure (Holland & Thayer, 1988), a method that is currently accepted as a valid procedure for identifying DIF. (See, for example, Hambleton & Rogers, 1989; Raju, Bode, & Larsen, 1989.) Because of the rather small samples, the score distribution was divided into seven roughly equal score categories to form the conditioning variable. The categories were based on the examination total score. Examinee group contrasts were the same as for the Rasch model method. Items were identified if their chi-squared statistic was significant at the .05 level. The items identified by the M-H procedure were compared to those identified by the Rasch analysis to demonstrate the comparability of these methods.

RESULTS

A description of the sample in terms of mean scores, standard deviations, and differences is presented in Table 1. Notice that the groups being compared did not perform equally well on the examination. The residency group had a mean score about half a standard deviation above that of the equivalency group. Also the Specialty I group had a mean score approximately a third of a standard deviation above that of the Specialty II group. For the DIF analyses, the higher scoring groups (residency and Specialty I) were considered the reference groups and the other two groups (equivalency and Specialty II), the focal groups. The groups were not independent, however, as Specialty I candidates were more likely to have followed the residency pathway and Specialty II candidates more often offered equivalency qualifications. Table 2 gives the joint frequencies.

Table 1
Sample Sizes, Mean and Standard Deviation of Scores, and Standardized Differences in Mean Scores Between Groups

Group	Number of Candidates	Raw Score		Difference between Groups in SD units[1]
		Mean	Standard Deviation	
Residency	225	169	25	.49
Equivalency	191	156	27	
Specialty I	145	168	28	.34
Specialty II	230	159	27	
Total	416	163	27	

[1]Difference between groups divided by the standard deviation of the total gorup.

Table 2
Joint Frequencies of Specialty and Pathway

Pathway	Specialty I	Specialty II
Residency	106	88
Equivalency	39	142

Analyses Based on Total Test

According to theory, if the groups being compared come from the same population, expected values of item parameters based on each group should be equal. However, if the groups have different training or experience, item parameters may have substantially different estimates across groups. Table 3 shows the number of items with inter-group differences of more than .50 in their difficulty with positive values indicating items favoring the focal group. Since the standard deviations of the difference distributions differed from one contrast to the other, a different number of items was identified for each contrast. Further, as expected from the procedure, the number of items favoring each group was roughly the same.

Table 3
Items with Large Intergroup Differences

Contrast	Number of Identified Items	Range of Differences (logits)
Residency-Equivalency	33	-.87 to .96
Specialty I - Specialty II	74	-1.14 to 2.46

Comparison with Mantel-Haenzsel Procedure

When an M-H procedure with gross ability blocking was used to identify DIF for the pathway contrast and the specialty contrast, fewer items were identified than with the Rasch procedure. Table 4 shows the number of items identified by each procedure and the number of items that were identified by both procedures.

The Rasch method identified 78 percent of the items identified by the M-H procedure for the pathway contrast and 87 percent for the specialty contrast. The most likely explanation for the difference in rates of agreement is that the effect sizes were generally larger for the specialty contrast. In every instance, the group that performed better on the item was the same according to both methods.

Table 4
Comparison of Items Identified by
Rasch and Mantel-Haenzsel Procedures

	Number of Items		
Contrast	Rasch	Mantel-Haenzsel	Both
Residency - Equivalency	33	23	18
Specialty I - Specialty II	74	60	52

Analysis by Content Area

Table 5 shows the number of items in each content area identified for each group. Inspection of these results showed clear patterns of each group's relative strength. In particular, the reference groups tended to perform better on items in Content Areas 2 and 5, areas that are important in the residency training, but less likely to be learned on the job if it isn't part of the specialty practiced. These areas are also more likely to be a part of Specialty I. The focal groups performed better in Area 4, which is more important in Specialty II. Areas 1 and 3, where more items favor the equivalency pathway, are areas that might be learned better in practice than in an academic setting. Area 6 would not be expected to favor any of the groups and relatively few items were identified from these areas.

Table 5
Number of Items Favoring Each Group in Each Content Area

Content	Residency	Equivalency	Specialty I	Specialty II
Area 1	0	4	4	5
Area 2	8	0	6	0
Area 3	1	4	5	9
Area 4	0	7	0	19
Area 5	7	1	21	1
Area 6	0	1	1	3

The most outstanding result is that no items in Area 2 favored either the equivalency or Specialty II medicine groups, and no Area 4 items favored either the residency or the Specialty I groups. In order to investigate the nature of the group differences on these two subtests, additional DIF analyses were performed using the Rasch procedure. One set of calibrations involved only the Area 2 items and the other involved only the Area 4 items of the core examination. This has the effect of forcing some items to appear to favor each group, even though differences on these items based on the total examination may be small. The numbers of items identified as exhibiting DIF for each group are presented in Table 6.

Table 6
Items in Subtests with Large Difficulty Differences Between Groups
Rasch Analysis

Contrast	No. of Identified Items	Range of Differences
Area 2		
Residency - Equivalency	5	-.74 to .59
Specialty I - Specialty II	6	-.77 to .57
Area 4		
Residency - Equivalency	4	-.75 to .56
Specialty I - Specialty II	10	-1.04 to 1.24

The differences in the range of DIF values suggests that both pathway and specialty are associated with the differences found in Area 2, but that the differences in Area 4 are more strongly associated with Specialty II. The items in these two content areas that were relatively easier for the disfavored group were found, on inspection, to cover more general material that might be expected to be known by non-specialists within the more general medical specialty.

DISCUSSION

Our initial hypothesis was that the results of the Rasch model analysis would be interpretable in terms of the known differences between the groups being compared. This hypothesis was supported when inspection of the content of the items identified appeared to be consistent with the known differences between the groups being compared. The primary patterns of content differences are summarized as follows:

1. Items favoring the residency group more often tested concepts that might be learned through academic study, whereas items favoring the equivalency group more often tested concepts likely to be learned in actual practice.

2. Items favoring each of the two specialty groups were found to relate to work settings corresponding to those most often held by practitioners in the respective specialty areas.

3. In the content areas with the clearest differences between groups, Areas 2 and 4, DIF analyses were performed separately to force identification of items favoring both groups. The resulting items found to "favor" the group less likely to be knowledgeable in those areas contained general material that might be known even by non-specialists in those areas.

The results were also reviewed by members of the specialty board that oversees the examination process. They confirmed that the results were sensible and in accordance with their expectations about differences among the groups. Combined with more specific findings than those reported here, they appeared to find the results an interesting and useful evaluation of their examination.

The agreement with the M-H results was substantial if less than perfect. Inspection of the items identified by the M-H method but not the Rasch showed that these items also conformed to the general patterns identified above. This suggests that the interpretation of the results would have been essentially the same regardless of the method used.

CONCLUSIONS

The results of this study support the validity of the Rasch model method for detecting DIF items. The items identified using this procedure agreed substantially with results using the M-H method, particularly for the contrast between specialties that had the larger effect sizes. Further, the results were found to be interpretable in terms of expected differences in background and training of the groups being compared. This interpretation was supported by experts in the medical specialty area being tested.

An implication of these results is that useful and interpretable results can be obtained without the use of more precise and elegant procedures. Although the more recently developed procedures might be preferred for other purposes, faulty methodology does not appear to be the source of uninterpretable results. Those researchers who may wish to use DIF analyses to help illuminate the pursuit of the original goals of improved test development and possibly fairer tests will need to look further than improvements in methodology.

REFERENCES

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias.* Baltimore: Johns Hopkins University Press.

Draba, R. E. (1977). *The identification and interpretation of item bias* (Research Memorandum No. 26). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2,* 313-334.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice, 8*(4), 5-11.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. *Applied Measurement in Education, 2,* 1-13.

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

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

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education, 1989, 2,* 255-275.

Shealy, R., & Stout, W. F. (1993). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 197-239). Hillsdale, NJ: Erlbaum.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias using both internal and external ability criteria. *Journal of Educational Statistics, 6,* 317-375.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Wright, B. D., & Linacre, J. M. (1992). *A user's guide to BIGSTEPS.* Chicago: MESA Press.

JOURNAL OF OUTCOME MEASUREMENT®, *2*(1), 43 - 65

Rasch Analysis of Distractors in Multiple-choice Items

Wen-chung Wang
National Chung Cheng University

In order to apply the Rasch model to multiple-choice items, incorrect responses to distractors are usually aggregated to a single category. In doing so, information of individual distractors disappears. In this paper, a Rasch-type analysis is proposed where one parameter is assigned to each distractor. The information is thus preserved. The proposed distractor model can be applied to investigate the performance of distractors, which is useful for item revision. This model is a necessary condition of the Rasch model, that is, fitting the distractor model will fit the Rasch model, but not vice versa. The results of a small simulation study show that parameter recovery of the distractor model is very satisfactory. A real data set of twenty multiple-choice items was analyzed. Some items were found to fit the Rasch model rather than the distractor model. It is this diagnostic value that makes the distractor model suitable for multiple-choice items.

Requests for reprints should be sent to Wen-Chung Wang, Department of Psychology, National Chung Cheng University, Chia-Yi, Taiwan.

It is widely recognized that when test items fit the Rasch model (Rasch, 1960), ability estimates and item difficulty estimates are mutually independent. The estimates are specifically objective (Rasch, 1960, 1967, 1968). In addition, the derived scale is interval (Wright & Masters, 1982; Wright & Stone, 1979). Because of these favorable features, use of the Rasch model to examine data has been highly recommended. The Rasch model is developed for dichotomous items. As multiple-choice (MC) items are usually dichotomously scored, fitting the Rasch model thus becomes a desirable feature. If MC items do not fit the Rasch model, usually either the multi-parameter models (e.g., Birnbaum, 1968) are used or the test items are to be revised. It seems that fitting the Rasch model is one of the best criteria for constructing MC items.

To utilize the Rasch model, all responses to distractors in an MC item are treated as a whole, the incorrect answers. Item difficulties are then estimated. In doing so, information of individual distractors is invisible. Within the framework of classical test theory, effectiveness of the distractors is evaluated by the usual item analysis, such as differences of the percentages in selections between upper ability groups and lower ability groups. If the differences are too small or even negative, the distractors are to be revised or discarded. The same idea can be adopted within the framework of item response theory.

To investigate how examinees react to distractors, the usual item difficulty parameter should be partitioned into several parts: one for each distractor. In the Rasch model, an examinee's response to an item is a function of his or her ability θ and the item difficulty δ. More specifically, the log-odd is as follows

$$log\,(P_1 / P_0) = \theta - \delta, \tag{1}$$

where P_1 is probability of a correct answer (scored as 1) of the examinee to that item; P_0 is that of an incorrect answer (scored as 0).

The responses to the item are dichotomously scored, either right or wrong. Applying the Rasch model to MC items, we usually treat all incorrect responses as a unified category. In doing so, effectiveness of individual distractors cannot be assessed. To assess the effectiveness, additional parameters should be incorporated. In this paper, a Rasch-type model for distractors in MC items is proposed. Implications and applications are addressed. The results of parameter recovery of a small simulation study are shown. Finally, an example of real data analysis is given.

THE DISTRACTOR MODEL

As shown in Equation 1, the log-odd of a correct response to an incorrect one is equal to ability minus item difficulty. Now, for MC items, there is more than one kind of incorrect responses, one for each distractor. For illustrative simplicity, consider a four-choice item. Let choice A be the correct choice and choices B, C, and D be the three distractors. Let P_A be the probability of the correct response, and P_B, P_C , and P_D be those of the incorrect responses to distractors B, C, and D, respectively. We can extend Equation 1 into the following three equations:

$$log\ (P_A / P_B) = \theta - \delta_B, \qquad (2a)$$
$$log\ (P_A / P_C) = \theta - \delta_C, \qquad (2b)$$
$$log\ (P_A / P_D) = \theta - \delta_D, \qquad (2c)$$

one for each distractor, respectively.

All the three above equations follow the Rasch model, as they are in the form of Equation 1. Figure 1a shows three item characteristic curves (ICCs) for the three equations, where δ_B, δ_C, and δ_D are set to be -1, 0, and 1, respectively, together with a curve for

$$log\ [P_A / (P_B+P_C+P_D)] = \theta - \delta\ , \qquad (2d)$$

where all the probabilities of the incorrect responses are summed to be the denominator. Equation 2d and Equation 1 are in fact equivalent because P_A is equivalent to P_1; $P_B+P_C+P_D$ is equivalent to P_0.

The four curves in Figure 1a never cross, a typical feature of the Rasch model. The four locations in the theta (ability) scale where the conditional probabilities are equal to .5 are δ_B, δ_C, δ_D, and δ (the usual item difficulty), from left to right. Figure 1b shows the curves of the unconditional probabilities. The intersection of the P_B curve and the P_A curve on the theta scale is δ_B. Likewise, that of the P_C curve and the P_A curve is δ_C; that of the P_D curve and the P_A curve is δ_D; that of the $P_B + P_C + P_D$ curve and the P_A curve is δ.

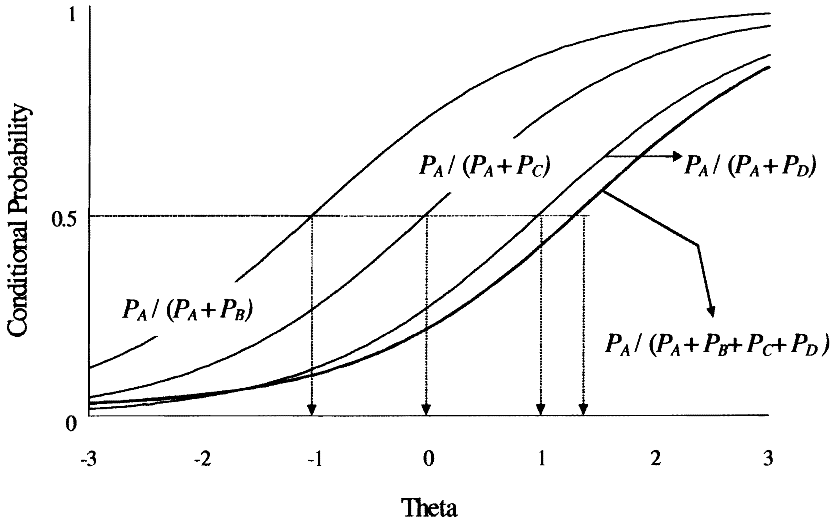

FIGURE 1a Conditional probabilities of a four-choice item when the distractor model is fitted.

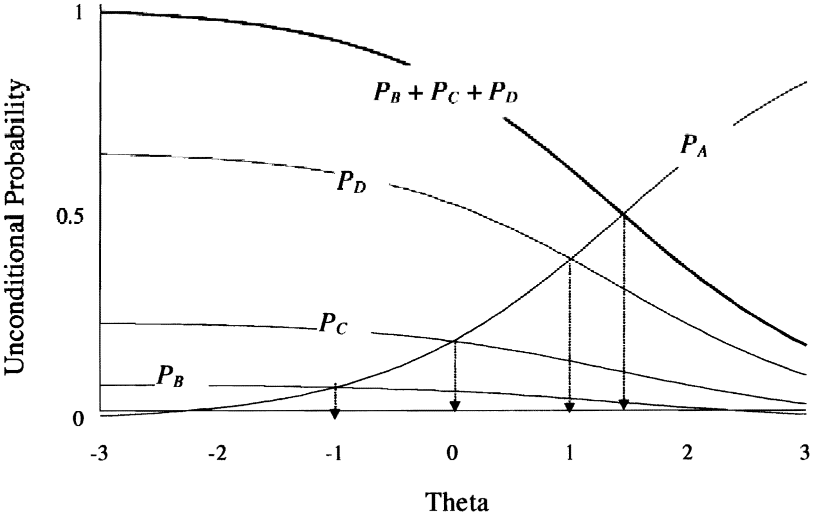

FIGURE 1b Unconditional probabilities of a four-choice item when the distractor model is fitted.

Given four response categories, it follows

$$P_A + P_B + P_C + P_D = 1.$$

From Equations 2a, 2b, and 2c, it can be shown that

$$P_B = P_A e^{\delta_B - \theta},$$
$$P_C = P_A e^{\delta_C - \theta},$$
$$P_D = P_A e^{\delta_D - \theta}.$$

Then,

$$P_A + P_A e^{\delta_B - \theta} + P_A e^{\delta_C - \theta} + P_A e^{\delta_D - \theta} = 1,$$
$$P_A \left(1 + e^{\delta_B - \theta} + e^{\delta_C - \theta} + e^{\delta_D - \theta}\right) = 1,$$

$$P_A = \frac{1}{1 + e^{\delta_B - \theta} + e^{\delta_C - \theta} + e^{\delta_D - \theta}}$$
$$= \frac{e^{\theta}}{e^{\theta} + e^{\delta_B} + e^{\delta_C} + e^{\delta_D}}.$$

P_B, P_C, and P_D can be derived in a similar way. In summary,

$$P_A = e^{\theta} / \Psi, \qquad (3a)$$
$$P_B = e^{\delta_B} / \Psi, \qquad (3b)$$
$$P_C = e^{\delta_C} / \Psi, \qquad (3c)$$
$$P_D = e^{\delta_D} / \Psi, \qquad (3d)$$

where Ψ is the sum of the four numerators.

To form a general expression of the proposed model, let there be M_i +1 choices in item i, indexed as $h = 0, \ldots, M_i$. Let the first choice be a correct answer and the others be distractors. Let b_{ih} denote the score of the response to choice h of item i. In this case,

$$b_{i0} = 1,$$
$$b_{ih} = 0, \text{ for h} = 1, \dots, M_i.$$

The probability of being in choice h of item i, given ability θ, is modeled as

$$P_{ih}(\theta) = \frac{e^{(b_{ih}\theta + \delta_{ih})}}{\sum_{k=0}^{M_i} e^{(b_{ik}\theta + \delta_{ik})}}.$$

For model identification, $\delta_{ih} \equiv 0$. This model is referred to as the *distractor model* hereafter, a Rasch-type model for distractors in MC items. Note that the distractor model is not limited to "real" distractors. The "omit" response or the "don't know" response can be viewed as distractors and analyzed accordingly.

If all incorrect responses are aggregated to a single incorrect category and the Rasch model is applied, only the item difficulty δ is estimated. If they are not aggregated and the distractor model is applied, the distractor parameters are estimated. In fact, once these parameters are estimated, the item difficulty can be calculated as follows. As shown in Figure 1a, δ is on the point where

$$P_A /(P_A + P_B + P_C + P_D) = .5.$$

Since the denominator of the above equation is 1 by definition, δ is then at the point where $P_A = .5$. From Equation 3a, it follows

$$e^{\theta} / \left(e^{\theta} + e^{\delta_B} + e^{\delta_C} + e^{\delta_D}\right) = .5.$$

Hence,

$$e^{\theta} = e^{\delta_B} + e^{\delta_C} + e^{\delta_D},$$

and

$$\theta = \delta = log\left(e^{\delta_B} + e^{\delta_C} + e^{\delta_D}\right).$$

As a general expression, the difficulty of item *i* is

$$\delta_i = log\left(\sum_{h=1}^{M_i} e^{\delta_{ih}}\right). \tag{4}$$

For the particular item shown in Figure 1 where $\delta_B = -1$, $\delta_C = 0$, and $\delta_D = 1$, it follows

$$\delta = log\left(e^{-1} + e^{0} + e^{1}\right).$$

Hence

$$\delta = 1.41,$$

which is identical to the location where P_A is .5 in Figure 1. Note also that the item difficulty is always greater than the three parameters. While the item difficulty can be obtained from the three parameters, the three parameters cannot be derived from the item difficulty. There can be infinite combinations.

The parameters in the distractor model are called *distractibility parameters*, rather than difficulty parameters. From Figure 1b, it can be easily found that for a given θ, the probability of selecting distractor *D* is greater than that of distractor *C*, which in turn is greater than that of distractor *B*. In other words, the larger the parameter is, the larger is the probability of selecting the corresponding distractor, and thus the more distractibility the distractor has.

From Equations 3a to 3d, it follows

$$log\ (P_C / P_B) = \delta_C - \delta_B,$$
$$log\ (P_D / P_B) = \delta_D - \delta_B,$$
$$log\ (P_D / P_C) = \delta_D - \delta_C.$$

For this particular item, the three log-odds are 1, 2, and 1, respectively. Therefore, the distractibility of distractor D is 1 logit higher than that of distractor C; which in turn is 1 logit higher than that of distractor B. With this information, the underlying structure of the distractors can be better clarified.

Potentially each item may tap a unique dimension, however, all items in a test are usually designed to measure a common dimension. This is because there are not so many substantial dimensions to be measured in a test. Moreover, use of raw scores as indices of examinees' ability leads to a single dimension. Unexceptionally, all MC items in a test are usually designed to measure a single dimension. This unidimensionality should hold not only between items but also within items (i.e., across choices). This is why the θs in Equations 2a to 2d are constrained to be identical. For discussions of between-item and within-item multidimensionality, see Wang, Wilson, & Adams (1997).

Suppose MC items do not fit the distractor model, there are potentially multiple dimensions within items. In such a case, the raw score loses its values on measurement, even though the MC items fit the Rasch model. Conversely, if the MC items fit the distractor model, the unidimensionality holds both between items and within items. Therefore, fitting the distractor model is more stringent for assessing MC items than fitting the Rasch model.

From Figures 1a or 1b, we find if data fit the distractor model (the P_A, P_B, P_C, and P_D curves), they will certainly fit the Rasch model (the P_A and the $P_B + P_C + P_D$ curves), because the $P_B + P_C + P_D$ curve actually comes from the sum of its individual curves. On the contrary, fitting the Rasch model does not imply fitting the distractor model. Figures 2a and 2b show one of the many possible cases where the Rasch model is fitted but the distractor model is not.

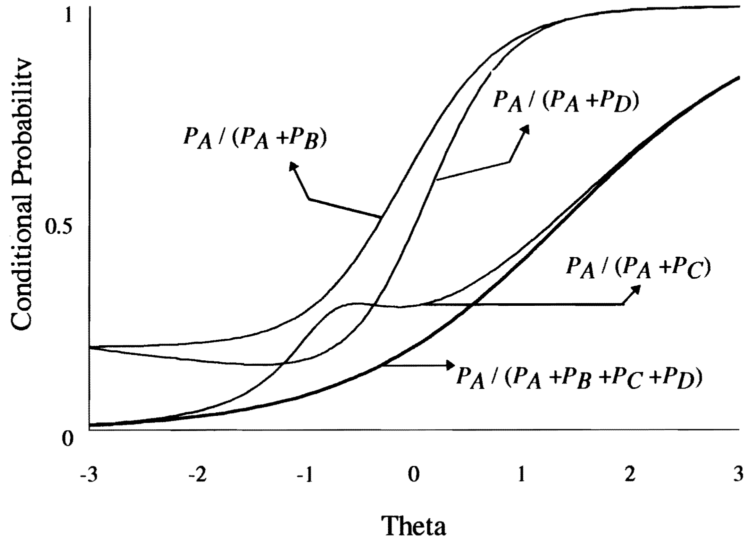

FIGURE 2a Conditional probabilities of a four-choice item when the distrator model is not fitted.

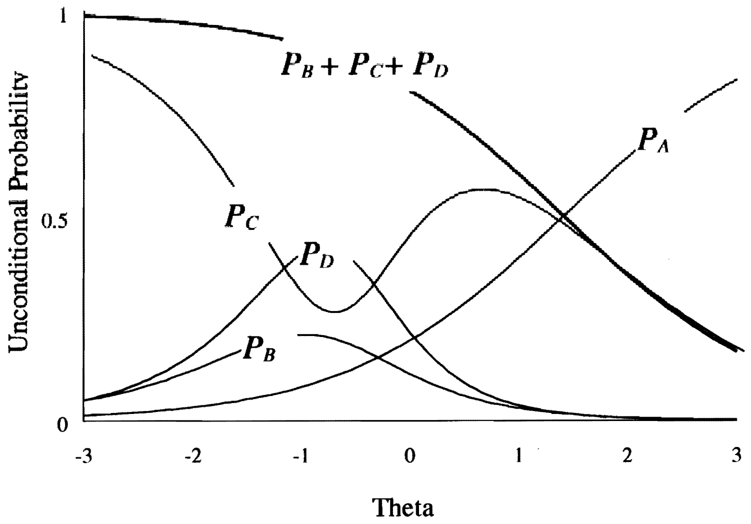

FIGURE 2b Unconditional probabilities of a four-choice item when the distrator model is not fitted.

ESTIMATION OF THE DISTRACTOR MODEL

The distractor model is a special case of the ordered partition model (OPM, Wilson, 1992; Wilson & Adams, 1993). The OPM is suitable for polytomous items where some response categories are given identical scores. Instead of aggregating the responses of the same scores into a unified category and applying the partial credit model (Masters, 1982), the original response categories are reserved and the OPM can be applied. Along with the same logic, original responses to distractors in MC items are reserved and the distractor model is used.

The OPM is in turn a special case of the multidimensional random coefficients multinomial logit model (MRCML, Adams & Wilson, 1996; Adams, Wilson, & Wang, 1997). The parameter estimation in this study is based on the MRCML and its corresponding computer software MATS (Wu, Adams, & Wilson, 1995). The software provides a marginal maximum likelihood estimation with EM algorithm (Bock & Aitkin, 1981). The MRCML is characterized by a scoring matrix and a design matrix. By manipulating the two matrices, many models including the OPM and the distractor model, of course, can be formed. The readers are referred to those above papers, Wilson & Wang (1995), Wang (1997), Wang & Wilson (1996), and Wang, Wilson, & Adams (1995) for details of implications and applications of the MRCML.

The distractor model is also a special case of the multiple-choice model proposed by Thissen & Steinberg (1984). The scoring function b_{ih} in the distractor model is defined a priori, which corresponds to the usual requirement of the Rasch family. Conversely, the function becomes a discrimination parameter in the multiple-choice model, a typical feature of the two-parameter item response model. In practice, the computer program MULTILOG (Thissen, 1991) can also be applied to estimate the distractibility parameters of the distractor model.

THE SIMULATION STUDY

A small simulation study was conducted. Twenty four-choice items were generated from the distractor model with a sample size of 1906. Fifty replications were made. The generating values of the parameters, shown in the second to the fourth columns in Table 2, were derived from the real data analyses in the latter section. Figure 3 shows the parameter recovery. Generally speaking, the biases (= generating value - mean estimated values) of the distractibility parameters are very small, within -.017 and .044. No systematic patterns of the biases across the generating values are found. In summary, the parameter recovery is very satisfactory.

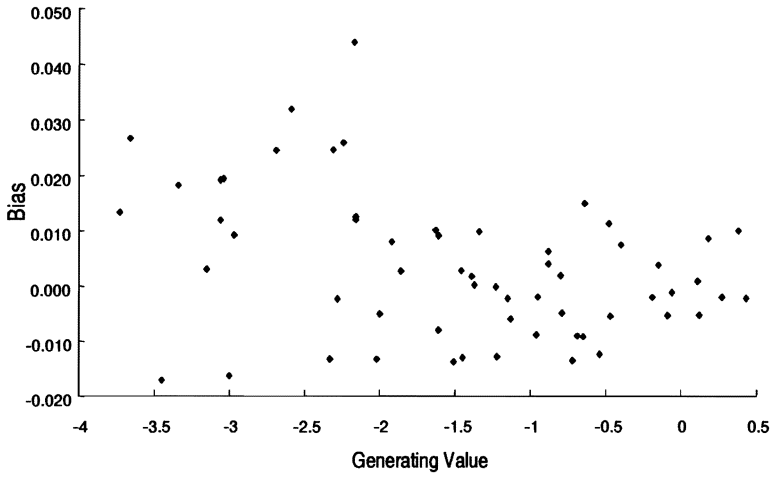

FIGURE 3 Generating values and biases of the simulation study.

A data set of 30 MC items generated from the distractor model was used to demonstrate that fitting the distractor model will also fit the Rasch model, but not vice versa. As expected, the data set fits the distractor model, as depicted from the mean square errors in Figure 4. Next, the incorrect responses were aggregated into one category. This category and the correct response category were analyzed by using the Rasch model. As shown in Figure 4, these aggregated items fit the Rasch model as well.

To see that fitting the Rasch model might not fit the distractor model, this data set was edited. The incorrect responses to the first three items were shuffled without changing the total number of incorrect responses. Figure 5 shows the mean square errors when the distractor model was applied. As expected, those of the first three items are far away from the expected value 1.0, indicating that the three items were flagged to be misfitting. Although this data set does not fit the distractor model, it fits the Rasch model very well. This is because the total numbers of the incorrect responses of all the items were not changed, thus, sufficient statistics for the parameters remain unchanged.

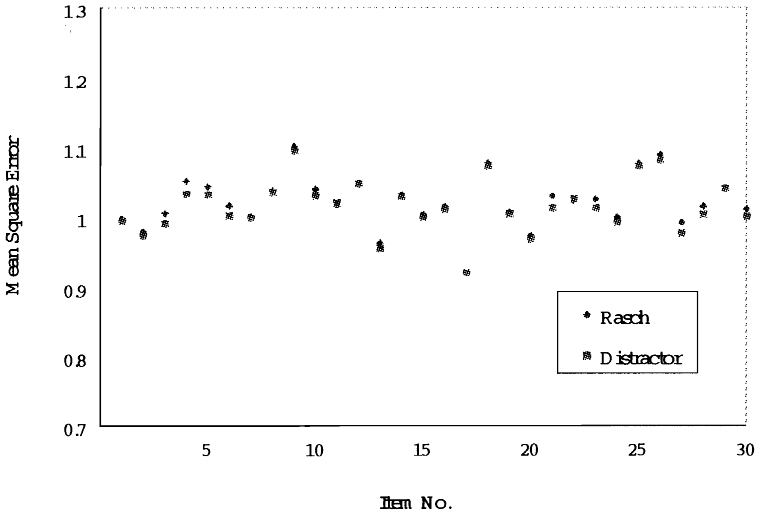

FIGURE 4 Mean square errors of 30 items when both the Rasch and the distractor models are fitted.

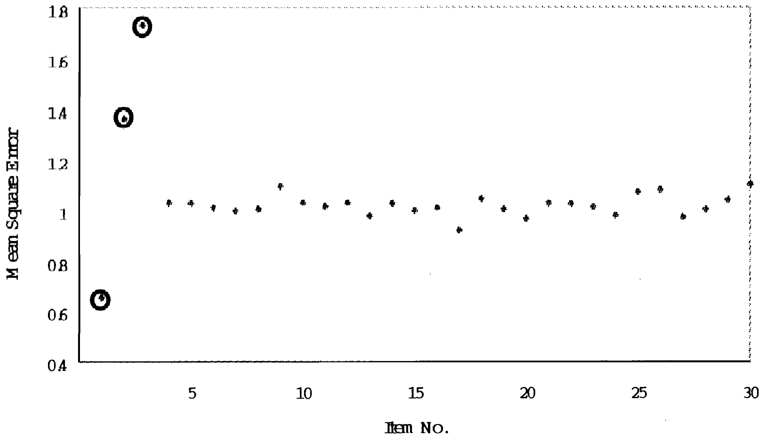

FIGURE 5 Mean square errors of 30 items where the first three items do not fit the distractor model.

REAL DATA ANALYSES

Twenty four-choice items from the biology subject of the 1995 Taiwan Joint College Entrance Examination were analyzed. The sample size is 1906. Table 1 shows percentages of responses to the four choices of the 20 MC items. For the first item, the percentages of responses to the three distractors are 48.0%, 12.3%, and 6.0%, respectively. With such large differences in the percentages, we expect that the distractibility of the three distractors would be very different. Rough estimates of the three distractibility parameters for the first item are .35 (= *log* (.48 / .34), -1.00 (= *log* (.12 / .34), and -1.72 (= *log* (.06 / .34), respectively. Distractor 1 has the highest distractibility and distractor 3 has the lowest distractibility among the three.

The second to the forth column in Table 2 show the distractibility parameters, based on the distractor model. The fifth column shows the calculated item difficulty, derived from Equation 4. The sixth column shows the estimated item difficulty, based on the Rasch model. The differences between these two kinds of item difficulties are very small, as shown in the last column. This is expected because they are based on the same sufficient statistics. The three distractibility parameters of the first item are .27, -1.13, and -1.86. These three parameters are close to the above three rough estimates. The first distractor is the most attractive and the third distractor is the least. The range of the three distractibility parameters is quite large, 2.13 (= .27 + 1.86). In general, we would like the distractibility of the distractors in the same item to be close. If one distractor is found considerably less attractive than others, test developers might revise the distractors or create new ones.

The three distractibility parameters for the last item are -2.31, -3.00, and -3.73. They are much smaller than those for the first item are. This does not imply that the three distractors are less attractive than those for the first item are. In fact, the distractibility can only be compared within items rather than across items, because the distractors are tied to the correct choice in the same item. After all, examinees select one choice from the choices given in the same item. The reason why the three distractibility parameters of the first item are larger than those of the last item are is partly because the first item is more difficult than the last item. Figures 6a to 6c depict the curvilinear relationships between the three distractibility parameters and the percentages of responses to the three distractors. In fact, the curve is exponential, which is similar to the relationship between item difficulties and frequencies of incorrect responses.

Table 1
Percentages of Responses to the Correct Choice and the Three Distractors

No.	Correct answer	Distractor 1	Distractor 2	Distractor 3	Valid case
1	33.7	48.0	12.3	6.0	1627
2	49.1	26.2	8.6	16.2	1613
3	31.2	19.4	29.0	20.4	1615
4	47.5	22.4	26.1	3.9	1628
5	42.8	22.4	24.5	10.3	1642
6	62.7	11.3	21.6	4.3	1632
7	28.5	17.5	47.7	6.3	1647
8	39.5	4.8	18.3	37.4	1597
9	52.0	6.1	15.5	26.4	1792
10	71.0	3.9	2.4	22.7	1703
11	26.6	30.6	31.0	11.8	1556
12	45.6	5.7	28.4	20.2	1576
13	27.9	5.7	24.1	42.3	1623
14	74.1	4.8	9.7	11.3	1693
15	62.9	2.9	8.1	26.1	1592
16	58.6	15.1	9.7	16.6	1601
17	59.6	19.0	15.5	5.9	1619
18	30.9	24.7	17.0	27.5	1621
19	84.7	4.0	5.8	5.6	1693
20	80.1	11.3	5.8	2.8	1692

Table 2
Distractibility Parameters, Calculated Difficulties, Estimated Difficulties, and Their Differences

No.	δ_B	δ_C	δ_D	Calculated difficulty	Estimated difficulty	Difference
1	0.27	-1.13	-1.86	0.58	0.58	0.001
2	-0.80	-1.92	-1.22	-0.11	-0.11	0.001
3	-0.47	-0.09	-0.48	0.77	0.77	0.001
4	-0.96	-0.79	-2.69	-0.10	-0.11	0.002
5	-0.72	-0.65	-1.45	0.22	0.22	0.000
6	-2.00	-1.34	-2.97	-0.80	-0.80	0.000
7	-0.54	0.43	-1.61	0.84	0.84	0.001
8	-2.24	-0.88	-0.15	0.33	0.33	0.001
9	-2.16	-1.23	-0.69	-0.10	-0.10	0.001
10	-3.15	-3.66	-1.37	-1.13	-1.13	0.001
11	0.18	0.11	-0.88	1.00	1.00	0.000
12	-2.17	-0.64	-0.95	0.03	0.03	0.000
13	-1.63	-0.19	0.38	0.91	0.91	0.000
14	-3.04	-2.28	-2.16	-1.33	-1.33	0.001
15	-3.34	-2.33	-1.15	-0.80	-0.80	0.000
16	-1.51	-2.02	-1.46	-0.54	-0.54	0.001
17	-1.39	-1.61	-2.59	-0.65	-0.65	0.001
18	-0.06	-0.40	0.12	1.01	1.01	0.001
19	-3.45	-3.06	-3.06	-2.08	-2.08	0.001
20	-2.31	-3.00	-3.73	-1.75	-1.76	0.002

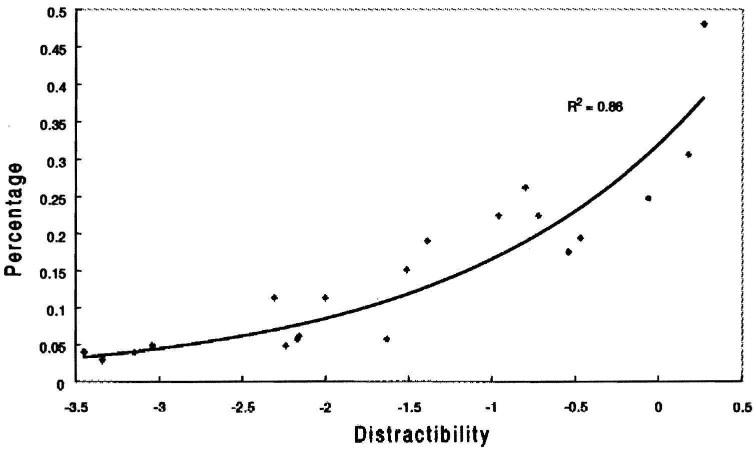

FIGURE 6a Distractor 1: relationship between parameter and percentage of responses.

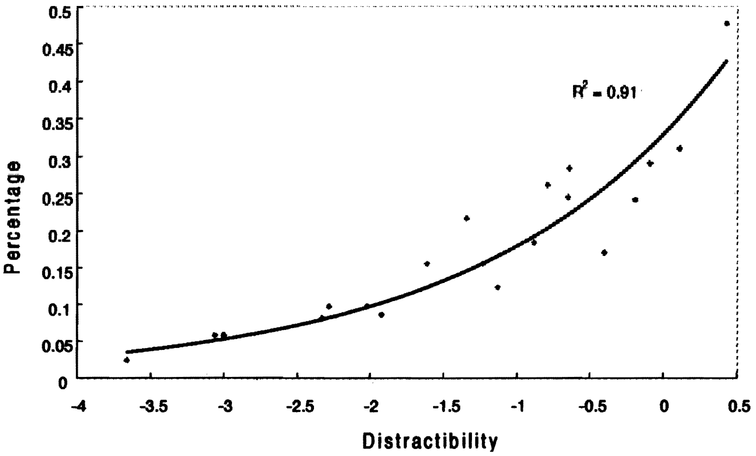

FIGURE 6b Distractor 2: relationship between parameter and percentage of responses.

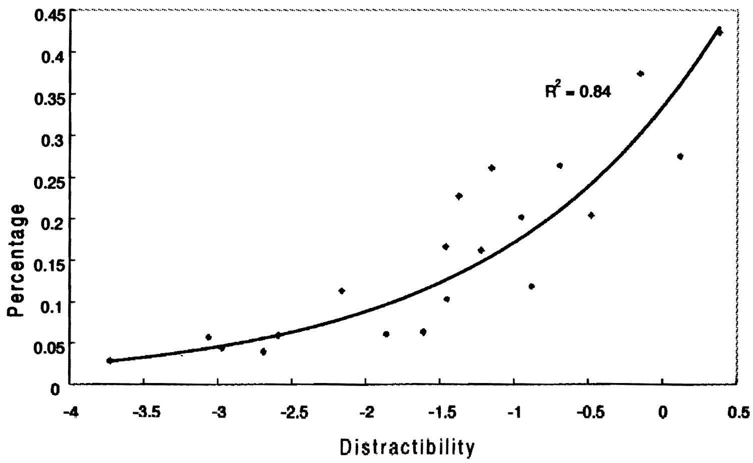

FIGURE 6c Distractor 3: relationship between parameter and percentage of responses.

Figures 7a to 7d show some examples that both the distractor model and the Rasch model are fitted. In the Rasch model, only the increasing ICC (the correct answer) is to be fitted. In the distractor model, the three decreasing ICCs are to be fitted. Since the increasing ICC is determined by the other three ICCs, if the latter are fitted, the former is fitted, too. In other words, once the distractor model is fitted, the Rasch model is fitted, too. Conversely, if the Rasch model is not fitted, the distractor model cannot be fitted, either.

Figures 8a and 8b show two examples where the Rasch model is fitted and the distractor model is not. More specifically, for item 11 (Figure 8b), the increasing ICC is fitted. However, the three decreasing ICCs are not. If only the Rasch model is applied, these two items cannot be identified as misfitting. Therefore, no item revision can be made. When the distractor model is applied, these items can be detected and further item revision can be done accordingly. This is the major advantage of the distractor model over the Rasch model. Finally, Figure 9 shows the extreme example where neither the distractor model nor the Rasch model is fitted.

Items that fit both the Rasch model and the distractor model.

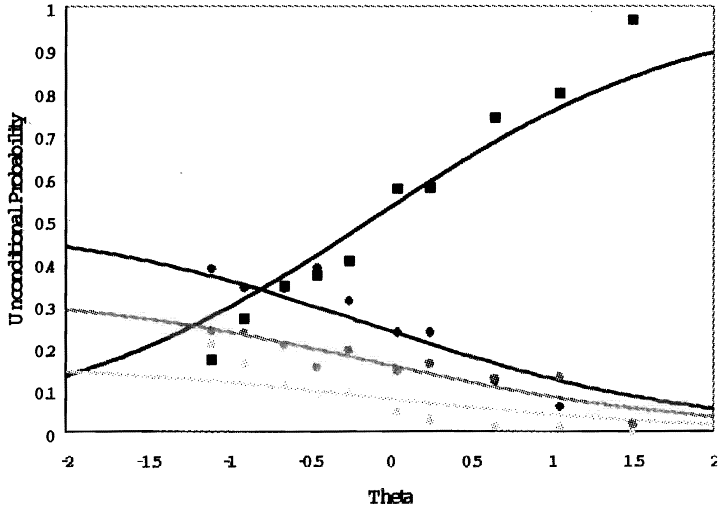

FIGURE 7a Item 2.

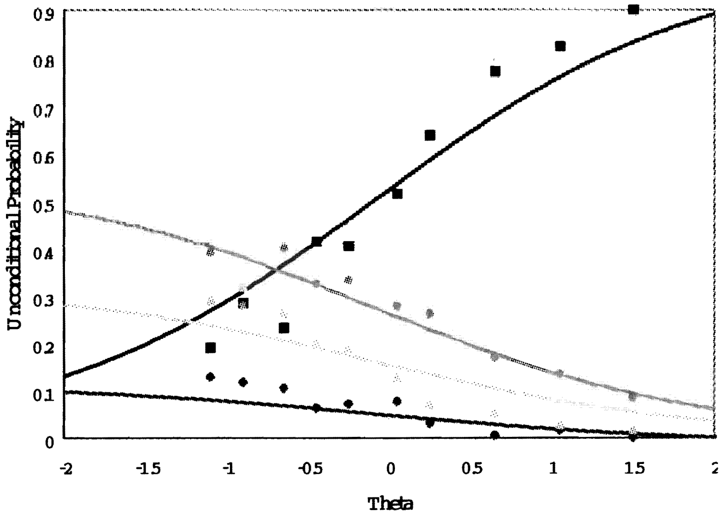

FIGURE 7b Item 9.

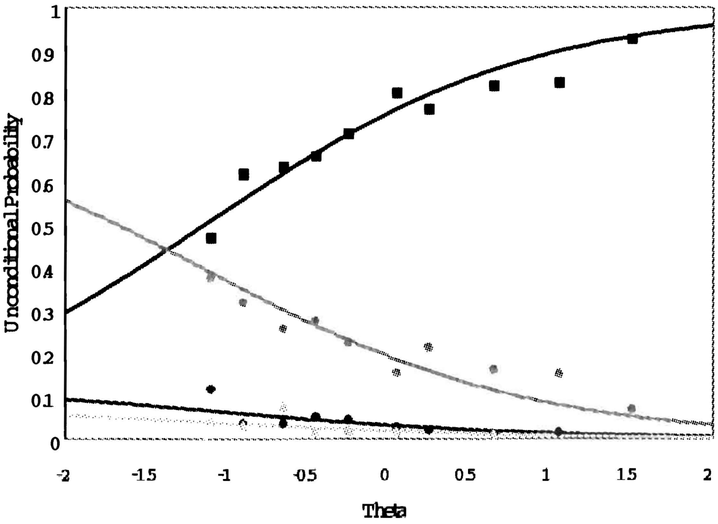

FIGURE 7c Item 10.

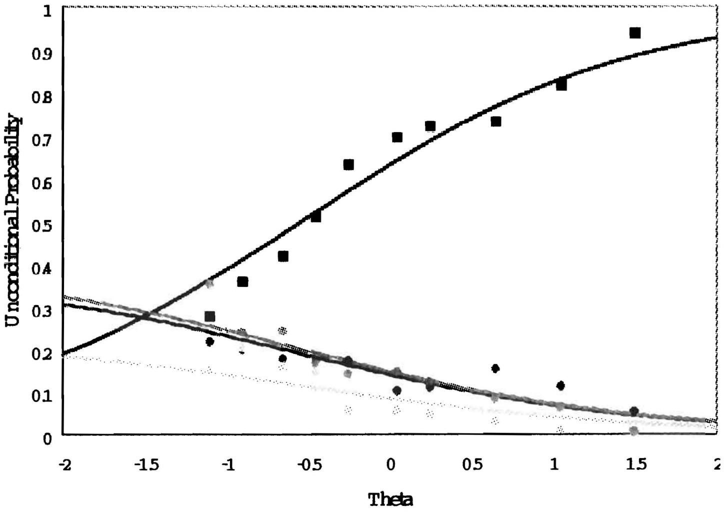

FIGURE 7d Item 16.

Items That Fit the Rasch Model But Not the Distractor Model

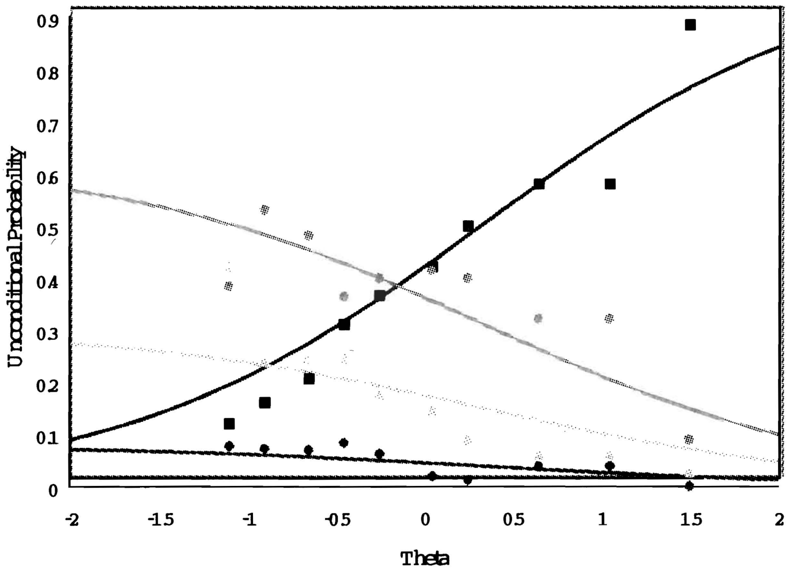

FIGURE 8a Item 8.

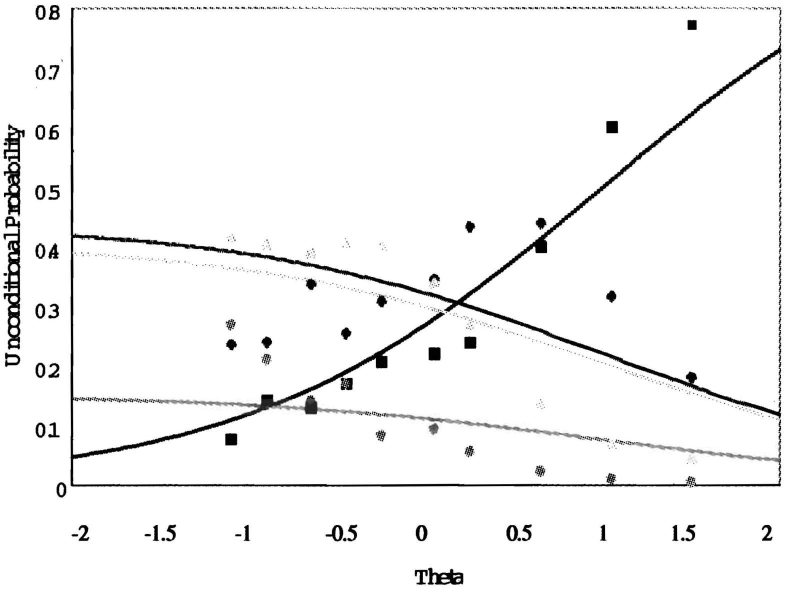

FIGURE 8b Item 11.

Item That Fits Neither the Rasch Model Nor the Distractor Model.

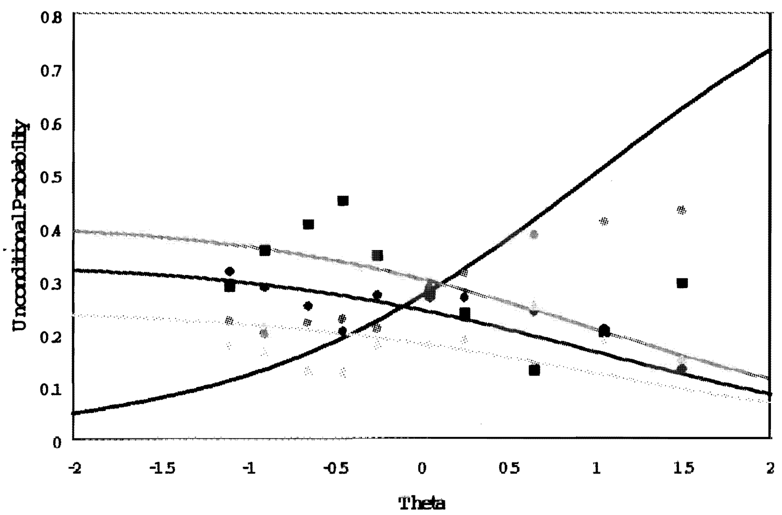

FIGURE 9 Item 18.

CONCLUSIONS

The Rasch dichotomous model may not be suitable for MC items because all incorrect responses are aggregated to a single category. Information of individual distractors disappears. Consequently, item revision becomes very difficult. In this paper, a Rasch-type model, the distractor model, is proposed. Being a member of the Rasch family, this model preserves the specific objectivity. Moreover, it can be applied to investigate distractibility of distractors in MC items by assigning one parameter to each distractor. When MC items fit the distractor model, they will also fit the Rasch model. Conversely, when they do not fit the Rasch model, they will not fit the distractor model, either. In other words, the distractor model is a necessary condition of the Rasch model. Therefore, fitting the distractor is more stringent for assessing MC items than fitting the Rasch model.

A small simulation study was conducted to show parameter recovery of distractibility parameters of the distractor model. The results show that the parameters are recovered very well. A real data set of 20 MC items from the 1995 Taiwan Joint College Entrance examination was analyzed by using both the Rasch model and the distractor model. Most items fit both models. Some items fit the Rasch model rather than the distractor model. Only a few items fit neither of the two models. For those fitting the Rasch model rather than the distractor model, item revision can be further done. It was found that some distractors are considerably less attractive than others. Test developers can use this information to revise the items or to investigate the underlying cognitive processes. It is the diagnostic feature that makes the distractor model especially useful for MC items.

Although the distractor model is more useful for assessing MC items than the Rasch model, a large sample size is needed for estimating distractibility parameters. If a distractor is rarely selected, its distractibility parameter will be inaccurately estimated, which in turn makes the information less valuable for item revision.

ACKNOWLEDGMENT

This study was supported by a research grant of National Chung Cheng University, Taiwan.

REFERENCES

Adams, R. J., & Wilson, M. R. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard and M. Wilson, (Eds.), *Objective measurement: Theory into practice*, Vol. 3. (pp. 143-166). Norwood, NJ: Ablex.

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika*, *46*, 443-459.

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

Rasch, G. (1960). *Probabilistic models for some intelligent and attainment tests.* Copenhagen: Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)

Rasch, G. (1967). An informal report and a theory of objectivity in comparisons. In L. J. Th. Van der Kamp & C. A. J. Vlek (Eds.), *Psychological measurement theory* (pp. 1-19). Proceedings of the NUFFIC international summer session in science at "Het Oude Hof", The Hague, July 14-28, 1966. Leyden: University of Leyden.

Rasch, G. (1968). *A mathematical theory of objectivity and its consequences for model construction.* Paper presented at the European Meeting on Statistics, Econmetrics and Management Science, Amsterdam, September 2-7,1968.

Thissen, D. (1991). *MULTILOG user's guide-version 6.* Chicago, IL: Scientific Software.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika, 49,* 501-519.

Wang, W. (1997). *Estimating rater severity with multilevel and multidimensional item response modeling.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Wang, W., & Wilson, M. R. (1996). Comparing multiple-choice-items and performance-based items using item response modeling. In G. Engelhard and M. Wilson, (Eds.), *Objective measurement: Theory into practice,* Vol. 3, (pp. 167-193). Norwood, NJ: Ablex.

Wang, W., Wilson, M. R., & Adams, R. J. (1995). *Item response modeling for multidimensional between-items and multidimensional within-items.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Wang, W., Wilson, M. R., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Theory into practice.* Vol. 4, (pp. 139-155). Norwood, NJ: Ablex.

Wilson, M. R. (1992). The partial order model: An extension of the partial credit model. *Applied Psychological Measurement, 16,* 309-325.

Wilson, M. R., & Adams, R. J. (1993). Marginal maximum likelihood estimation for the ordered partition model. *Journal of Educational Statistics, 18,* 69-90.

Wilson, M. R., & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19,* 51-72.

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

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

Wu, M., Adams, R. J., & Wilson. M. R. (1995). *MATS: Many-aspect test software.* Paper presented at the Annual Meeting of the American Educational -

JOURNAL OF OUTCOME MEASUREMENT®, 2(1), 66-78

Using Item Mean Squares to Evaluate Fit to the Rasch Model

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

Randall E. Schumacker
University of North Texas

M. Joan Bush
Irving Independent School District

Throughout the mid to late 1970's considerable research was conducted on the properties of Rasch fit mean squares. This work culminated in a variety of transformations to convert the mean squares into approximate t-statistics. This work was primarily motivated by the influence sample size has on the magnitude of the mean squares and the desire to have a single critical value that can generally be applied to most cases. In the late 1980's and the early 1990's the trend seems to have reversed, with numerous researchers using the untransformed fit mean squares as a means of testing fit to the Rasch measurement models. The principal motivation is cited as the influence sample size has on the sensitivity of the t-converted mean squares. The purpose of this paper is to present the historical development of these fit indices and the various transformations and to examine the impact of sample size on both the fit mean squares and the t-transformations of those mean squares. Because the sample size problem has little influence on the person mean square problem, due to the relatively short length (100 items or less), this paper focuses on the item fit mean squares, where it is common to find the statistics used with sample sizes ranging from 30 to 10,000.

An earlier version of this paper was presented at the 1995 Annual Meeting of the American Educational Research Association in San Francisco. Requests for reprints should be sent to Richard M. Smith, RFI, P.O. Box 675, Wheaton, IL, 60189.

Recent presentations at the Rasch Measurement SIG sessions at AERA have stressed the use of the weighted and unweighted item mean squares as a means to evaluate the fit of the responses to a Rasch model. This evaluation is usually based on a single critical value on the order of 1.2 to 1.3 for both mean squares. The rationale usually given is that the mean square is less affected by sample size than the approximate t-statistic resulting from the cube root transformation of the fit mean square. These arguments are contradictory to the arguments used in the late 1970's and early 1980's when the fit mean square transformations were developed.

HISTORY OF FIT

One of the methods of assessing fit in Rasch measurement models, and the technique that is used in most of the calibration and analysis programs distributed by MESA Press, is based on concatenation of the item/person residual. Other methods, such as those based on the likelihood ratio chi-square, will not be discussed in this paper. There is an approximately parallel history of development for item and person fit statistics based on the item/person residual (Smith, 1989 and Smith, 1991b); however, only the development of the item fit statistics, the object of inquiry in this study will be detailed here.

The item fit statistic, first proposed by Wright and Panchapakesan (1969), was based on person raw score groups which focused on the difference between the observed and expected score for a group of persons with the same raw score on a test. Subsequent developments in fit statistics have been based on the item/person residual. The unweighted item total fit statistic (UT) in the chi-square form, based on the item/person residual (y_{ni}) is

$$x^2(UT)_i = \sum_{n=1}^{N} y^2_{ni}. \tag{1}$$

The standardized residual y_{ni} is

$$y_{ni} = \frac{(x_{ni} - p_{ni})}{(w_{ni})^{1/2}}, \tag{2}$$

where x_{ni} is the observed score for each item/person interaction, p_{ni} is the probability of a correct response for each interaction, and $w_{ni} = p_{ni}(1 - p_{ni})$. This chi-square is calculated for each item by summing over all of the persons in the response matrix.

This chi-square can be converted to a mean square by dividing by the number of persons (N),

$$MS(UT)_i = \left(\frac{1}{\mathrm{N}}\right)\chi^2(UT)_i = \left(\frac{1}{N}\right)\sum_{n=1}^{N}\frac{(x_{ni} - p_{ni})^2}{w_{ni}}. \tag{3}$$

Note that the degrees of freedom used to convert these and subsequent total fit statistics to mean squares are N rather than the (N-1) used with the Wright Panchapakesan χ^2. This is due to the fact that the (N-1) overcorrects for the loss in degrees of freedom due to using the same x_{ni} to estimate the item and person parameters used in calculating the p_{ni} and to calculate the score residual. Alternative methods for correcting for the loss in degrees of freedom are discussed Smith (1982, 1991b).

The standard deviation of this mean square can be estimated by

$$s[MS(UT)_i] = \frac{\left[\sum_{i=1}^{N}\frac{1}{w_{ni}} - 4N\right]^{1/2}}{N}. \tag{4}$$

These statistics originally were evaluated as fit mean squares (FMS)in BICAL, an early Rasch calibration program. Where MS(UT) has an expected value of one and the standard deviation given in Equation 4. The critical values for detecting misfit with this mean square depend on the number of persons and $w_{ni,}$ so they will vary from item to item and sample to sample. To simplify the critical value problem, the mean square can be standardized to an approximate unit normal by a variety of transformations. This transformation, the unweighted total item fit statistic, is discussed in Wright and Stone (1979).

Later versions of BICAL introduced a log transformation in an attempt to standardize the fit statistics to an approximate unit normal distribution. In this transformation

$$t = \left[ln\left(MS(UT)_i\right) + MS(UT)_i - 1\right]\left[\frac{f}{8}\right]^{1/2}, \tag{5}$$

where f is (N-1) for the unweighted total item fit statistic. These transformations were introduced because the values of the mean squares which indicate possible misfit varied from item to item and analysis to analysis depending on the number of persons, the distribution of item difficulties, and the distribution of person abilities.

The last version of BICAL introduced a cube root transformation to convert MS(UT) to approximate unit normals. In this transformation

$$t=\left[\left(MS^{1/3}-1\right)\left(3/S\right)\right]+\left(S/3\right), \tag{6}$$

where S is the standard deviation of MS(UT) or MS(UB) given above in equation 4.

Experience with the unweighted fit statistic indicated that when there was a large range of item difficulties and person abilities, unexpected correct responses by low ability persons to difficult items and unexpected incorrect responses by high ability persons to easy items affected the unweighted mean square severely. A relatively small number of anomalous responses can result in unusually large mean squares and t-statistics.

The last version of BICAL also introduced the weighted version of the total item fit statistic, which replaced the unweighted version in that program. The weighted item total fit statistic was developed to diminish the effect of anomalous outliers. In this statistic the squared standardized residual (y_{ni}^2) is weighted by the information function (w_{ni}). The weighted item total fit statistic (WT) in the chi-square form is

$$MS(WT)_i=\frac{\sum_{n=1}^{N} w_{ni}\frac{\left(x_{ni}-p_{ni}\right)^2}{w_{ni}}}{\sum_{n=1}^{N} W_{ni}}=\frac{\sum_{n=1}^{N}\left(x_{ni}-p_{ni}\right)^2}{\sum_{n=1}^{N} w_{ni}}. \tag{7}$$

The weighted total mean square is the sum of the weighted squared standardized residuals divided by the sum of the weights. The standard deviation of this statistic is

$$s\left[MS\left(WT\right)_i\right]=\frac{\left[\sum_{n=1}^{N} w_{ni}-4\sum_{n=1}^{N} w^2{}_{ni}\right]^{1/2}}{\sum_{n=1}^{N} w_{ni}}. \tag{8}$$

The weighted version of the total fit statistic is less affected by anomalous responses by persons with ability far from the difficulty of the item. A further description of the weighted total fit statistic can be found in Wright and Masters (1982) and Smith (1991b).

In recent programs, e.g., BIGSCALE, BIGSTEPS, and FACETS, the unweighted fit statistics (item and person) have become known as OUTFIT statistics and the weighted fit statistics have become known as INFIT statistics.

This study was designed to illustrate the differences between the fit mean squares and the transformed version of the item fit statistics. This comparison focused on the use of a single critical value to determine misfit and effect of sample size and the type of statistic being evaluated (OUTFIT vs. INFIT) on the distribution of the item fit mean square. The Type I error rates of the fit mean square are then compared with those of the transformed t-statistic.

METHODS

In this study 100 replications of simulated data were generated under each of six different conditions which varied the number of persons and the number of items. These conditions were: 150 persons with 20 and 50 item tests, 500 persons with 20 and 50 item tests, and 1000 persons with 20 and 50 item tests. Person abilities were normal with a 0, 1 distribution. Item difficulties were uniformly distributed from -2.0 to + 2.0 logits (See Schumacker, Smith, and Bush (1994) for a complete description of the simulated data.). All simulated data sets were calibrated with the BIGSTEPS program (Wright and Linacre, 1991). For each calibration an item file was generated which contained the weighted and unweighted mean squares and t-statistics for each of the items in that calibration. The mean, standard deviation, minimum value, maximum value, and per cent of cases above given critical values were calculated for each of four statistics, weighted mean squares and t-statistics and unweighted mean squares and t-statistics, in each data set. These summary statistics were then averaged across the 100 replications in each combination of test length and number of persons. The critical values used to calculate the percent of cases with extreme values were fms>1.3, fms>1.2, fms>1.1, fms<.9, fms<.8, and fms<.7 for the mean squares and t>+4, t>+3, t>+2, t<-2, t<-3, and t<-4 for the t-statistics.

RESULTS

The results presented in the following tables are based on a summary of the 100 replications for each of the six conditions. The mean squares and t-statistics used in this analysis were obtained from the item file option available in the BIGSTEPS program. The summary information for the weighted mean squares is presented in Table 1 and in Table 2 for the unweighted mean squares.

The means of both mean squares (unweighted and weighted) are very stable about the expected value of 1.00. The average weighted mean square means have a standard deviation of 0.00 across the six conditions, and the unweighted mean square means have a maximum standard deviation of 0.03 across the six conditions. Thus, the number of persons and the length of the test appear to have a small influence on the mean of the unweighted mean squares (SD $\leq$.03), and the influence on the mean of the weighted mean squares cannot be seen in the second decimal point (all SD = 0.00).

The standard deviation of the mean squares varies considerably based on the type of mean square (weighted and unweighted) and the number of persons. The mean standard deviation for the unweighted mean squares is approximately double that of the weighted mean square. For example, the mean standard deviation for the unweighted mean square varies from 0.18 with 150 persons to 0.06 for 1000 persons (Table 2). The mean standard deviation of the weighted mean square varies from 0.08 for 150 persons to 0.03 for 1000 persons (Table 1). The standard deviation does not appear to be affected by the number of items, as seen by comparing the top and bottom halves of Tables 1 and 2.

The range of the mean squares is similarly affected. The mean range for the unweighted mean square is 0.72 for 150 persons and 20 items, 0.40 for 500 persons and 20 items, and 0.25 for 1000 persons and 20 items, but the number of items on the test has little effect on the range of the unweighted mean square. Contrast this with the range of the weighted mean square. In the same example given above, the mean range for the weighted mean square is 0.29 for 150 persons and 20 items, 0.16 for 500 persons and 20 items, and 0.10 for 1000 persons and 20 items. These are less than one-half of the range for the unweighted mean squares. As with the unweighted mean square, there appears to be considerable influence resulting from the number of persons and little influence resulting from test length on the range of the mean squares.

To examine the Type I error rates and the influence of mean square type, number of persons and test length, six critical values were chosen and the percentage of mean squares exceeding those values were calculated. These results are presented in Table 3. Values greater than 1.2, a

Table 1
Weighted Mean Square Descriptive Statistics

Simulation 1 (150 persons, 20 items)

SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	0.99	1.01
S.D.	0.08	1.01	0.05	0.11
Maximum	1.15	0.04	1.08	1.33
Minimum	0.86	0.03	0.81	0.92

Simulation 2 (500 persons, 20 items)

SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	0.99	1.00
S.D.	0.04	0.01	0.03	0.06
Maximum	1.08	0.02	1.04	1.15
Minimum	0.92	0.02	0.87	0.96

Simulation 3 (1000 persons, 20 items)

SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	0.99	1.00
S.D.	0.03	0.00	0.02	0.04
Maximum	1.05	0.01	1.03	1.09
Minimum	0.95	0.01	0.91	0.97

Simulation 4 (150 persons, 50 items)

SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	0.99	1.00
S.D.	0.07	0.01	0.05	0.09
Maximum	1.16	0.03	1.09	1.25
Minimum	0.85	0.03	0.75	0.91

Simulation 5 (500 persons, 50 items)

SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	1.00	1.00
S.D.	0.04	0.00	0.03	0.05
Maximum	1.09	0.02	1.05	1.16
Minimum	0.92	0.02	0.86	0.95

Simulation 6 (1000 persons, 50 items)

SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	1.00	1.00
S.D.	0.03	0.00	0.02	0.03
Maximum	1.06	0.01	1.04	1.11
Minimum	0.94	0.01	0.91	0.96

Table 2
Unweighted Mean Square Descriptive Statistics

Simulation 1 (150 persons, 20 items)

SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.03	0.95	1.11
S.D.	0.18	0.07	0.10	0.53
Maximum	1.45	0.31	1.17	3.17
Minimum	0.73	0.06	0.58	0.86

Simulation 2 (500 persons, 20 items)

SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.01	0.97	1.04
S.D.	0.10	0.03	0.05	0.19
Maximum	1.25	0.13	1.05	1.80
Minimum	0.85	0.04	0.73	0.91

Simulation 3 (1000 persons, 20 items)

SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.01	0.98	1.02
S.D.	0.06	0.01	0.03	0.10
Maximum	1.14	0.06	1.03	1.34
Minimum	0.89	0.03	0.80	0.95

Simulation 4 (150 persons, 50 items)

SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.01	0.98	1.05
S.D.	0.16	0.03	0.11	0.35
Maximum	1.52	0.28	1.17	3.19
Minimum	0.71	0.05	0.58	0.81

Simulation 5 (500 persons, 50 items)

SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.01	0.98	1.02
S.D.	0.08	0.02	0.06	0.16
Maximum	1.25	0.14	1.10	1.99
Minimum	0.82	0.04	0.72	0.88

Simulation 6 (1000 persons, 50 items)

SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	0.99	1.01
S.D.	0.06	0.01	0.04	0.09
Maximum	1.18	0.07	1.09	1.51
Minimum	0.87	0.03	0.79	0.92

Table 3
Mean Square Frequency of Extreme Values

Weighted	Simulation Conditions*					
	1	2	3	4	5	6
% > 1.3	0.05	0.00	0.00	0.00	0.00	0.00
% > 1.2	0.60	0.00	0.00	0.32	0.00	0.00
% > 1.1	8.05	0.60	0.00	6.90	0.40	0.04
%< 0.9	8.35	0.65	0.00	6.62	0.20	0.00
%< 0.8	0.00	0.00	0.00	0.08	0.00	0.00
%< 0.7	0.00	0.00	0.00	0.00	0.00	0.00

Unweighted	Simulation Conditions					
	1	2	3	4	5	6
% > 1.3	4.75	1.35	0.05	3.48	0.52	0.12
% > 1.2	10.05	3.90	0.65	8.44	1.76	0.48
% > 1.1	21.40	12.50	4.85	20.70	8.72	4.38
%< 0.9	28.05	11.15	3.10	23.56	8.88	2.70
%< 0.8	8.30	0.50	0.00	6.36	0.48	0.04
%< 0.7	1.30	0.00	0.00	0.84	0.00	0.00
N of Persons	150	500	1000	150	500	1000
N of Items	20	20	20	50	50	50

* Each simulation condition contained 100 replications.

commonly used value for detecting measurement disturbances, occurred less than 1 time per 200 for all sample sizes and test lengths for the weighted mean square and values greater than 1.1 occurred less than 1 time per 200 for sample sizes greater than 500. If weighted mean square critical value of 1.2 were to be used, then the Type I error rate would approximate .005. With the weighted mean squares the per cent greater than the critical value is too small in most cases to accurately judge the effect of test length on the statistic.

For the unweighted mean square and sample size of 150, values greater than 1.3 occurred at a rate of approximately 5 per cent. For sample size of 500, values greater than 1.3 occurred at a rate of approximately 1 per cent. For sample size of 1000, values greater than 1.3 occurred at a rate of approximately .1 per cent. To have a consistent Type I error rate of approximately .05, a critical value of 1.3 would be needed with 150 person samples, 1.2 with 500 person samples, and 1.1 with 1000 person samples. It is also clear from these data that unweighted mean square is moderately affected by test length with the per cent above the critical value approximately 1 per cent higher for the 20 item tests than for the 50 item tests.

It is also clear from the values listed in Table 3 that the mean square is not symmetrically distributed about 1.0. Extreme values occur far less frequently below 1.0 then above. This means that symmetrical critical values for detecting misfit would operate at different Type I error rates for the upper and lower tails of the distribution.

The results of these simulations suggest that no single critical value will work with both weighted and unweighted mean squares. It is also clear that no single value will work with different sample sizes. If a critical value of 1.2 were chosen, the actual Type I error rate could vary anywhere from 0.00001 to 0.10 depending on the set of circumstances.

In an effort to contrast the use of the mean square with the transformed t-statistic, the frequency of extreme values for the same simulations were calculated. These are presented in Table 4. In this table the critical values chosen were +4, +3, +2, -2, -3, and -4. There is no equivalence implied between these values and the values chosen for use in Table 3. They are simple convenient numerical values. The +2.0 value is often used as an indication of misfit with the t-statistic. As is clear from this table, the Type I error rate for the unweighted t-statistic is approximately twice the value for the weighted version. However, the differences across the weighted and unweighted version of the t-statistic are less extreme then across the two versions of the mean square. For example for 150 persons and 20

Table 4
t-statistic Frequency of Extreme Values

Weighted						
	Simulation Conditions*					
	1	2	3	4	5	6
% > 4.0	0.05	0.00	0.00	0.00	0.00	0.00
% > 3.0	0.10	0.00	0.00	0.02	0.04	0.08
% > 2.0	1.35	0.60	0.40	1.02	0.66	0.64
%< -2.0	0.65	1.70	1.10	0.70	0.70	0.90
%< -3.0	0.00	0.05	0.05	0.06	0.00	0.04
%< -4.0	0.00	0.00	0.00	0.00	0.00	0.00
Unweighted						
	Simulation Conditions					
	1	2	3	4	5	6
% > 4.0	0.10	0.05	0.00	0.08	0.10	0.06
% > 3.0	0.40	0.50	0.10	0.24	0.22	0.26
% > 2.0	2.60	2.45	1.40	2.24	1.80	1.56
%< -2.0	0.35	1.25	0.90	0.40	0.62	0.80
%< -3.0	0.00	0.00	0.00	0.02	0.00	0.00
%< -4.0	0.00	0.00	0.00	0.00	0.00	0.00
N of Persons	150	500	1000	150	500	1000
N of Items	20	20	20	50	50	50

* Each simulation condition contained 100 replications.

items the Type I error rate for the unweighted t-statistic value of +2.0 is 0.026 for the weighted t-statistic the value is 0.0135, approximately a two-fold difference. For the mean square with 150 persons and 20 items the Type I error rate for a value of 1.2 is 0.006 for the weighted version and 0.10 for the unweighted version, approximately a 15-fold difference. This difference is far greater than with the t-statistic. Also, the differences across sample size are less drastic with the t-statistic then with the mean square. The Type I error rates for the unweighted t-statistic critical value of +2.0 with 150 and 1000 persons are 0.026 and 0.014, a multiple of about 2. The Type I error rates for the unweighted mean square critical value of 1.2 with 150 and 1000 persons are 0.10 and 0.0065, a multiple of about 15.

Although it is clear from these simulations that the use of a single critical value for the t-statistic may lead to different Type I error rates for different statistics, sample sizes and test lengths, the effect of these three factors on the statistics is less than those observed for the mean squares. It should be noted that Smith (1982, and 1991b) has proposed several methods for removing the differences found across fit statistics due to the differences in sample size and type of statistic. The values reported in this study were generated by BIGSTEPS which does not employ these corrections. If these corrections were employed, the dissimilarity between the Type I error rates for the t-statistics would be less than those observed here.

DISCUSSION

Clearly these results indicate that the critical value for the mean square used to detect misfit is affected both by the type of the mean square and the number of persons in the calibration. A single critical value, particularly one of 1.2 or 1.3 will not give a .05 Type I error rates for sample sizes of 500 or larger. For the weighted version (INFIT) even a value of 1.1 is too large for sample sizes more than 500. These results have serious implications for BIGSTEPS users since the item fit mean squares have become the preferred method with which the fit of the data to the model is determined. Many authors suggest that the mean square is less sensitive to large sample size that the t-transformation. These results show that this is not the case. The mean squares are more sensitive to sample size and reliance on a single critical value for the mean square can result in an under detection of misfit.

Wright (personal communication, 1996) suggests that the critical value for mean squares might be calculated by the following formulas which indicates the direct influence of sample size on the two mean squres,

critical value $MS(WT)=1+\frac{2}{\sqrt{x}}$, and

critical value $MS(UT)=1+\frac{6}{\sqrt{x}}$,

where x=the sample size. This formula would yield critical values of 1.16 for sample sizes of approximately 150, 1,09 for for sample sizes of approximately 500, and 1.06 for sample sizes of approximately 1000 for the weighted mean square. Critical values for the unwieghted mean square would be 1.48 for 1.27 for sample sizes of approximately 150, sample sizes of approximately 500, and 1.19 for sample sizes of approximately 1000. Further research is needed to establish the exact Type I error rate for these approximate critical values and to examine the impact of larger sample sizes (n=1500 and n=2000) on the data.

REFERENCES

Schumacker, R. E., Smith, R. M., Bush, M. J. (1994). Examining replication effects in Rasch fit statistics. A paper presented at the 1994 annual meeting of the American Educational Research Association.

Smith, R. M. (1982). Detecting measurement disturbances with the Rasch model. Unpublished doctoral dissertation. University of Chicago.

Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological Measurement, 48*, 657-667.

Smith, R. M. (1989). Item and person fit in the Rasch model. A paper presented at the 1989 annual meeting of the American Educational Research Association.

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

Smith, R. M. (1991b). *IPARM: Item and person analysis with the Rasch model.* Chicago: MESA Press.

Smith, R. M. (1992). *Applications ofRasch measurement.* Chicago: MESA Press.

Wright, B. D. and Linacre, J.M. (1991). *BIGSTEPS: Rasch analysis for all two-facet models.* Chicago: MESA Press.

Wright, B. D. and Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

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

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

CONTRIBUTOR INFORMATION

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

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

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

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

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

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

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

SUBSCRIBER INFORMATION

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