JOURNAL OF OUTCOME MEASUREMENT®

Volume 2, Number 4 1998

Indexing/Abstracting Services: JOM is currently indexed in the *Current Index to Journals in Education* (ERIC), *Index Medicus*, and MEDLINE. The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Reviewer Acknowledgement

The Editor would like to thank the following people who provided manuscript reviews for the Journal of Outcome Measurement, Volume 2.

David Andrich, *Murdoch University, Australia*
Betty Bergstrom, *CAT Inc.*
Trevor Bond, *James Cook University, Australia*
Ayres D'Costa, *Ohio State University*
Barbara Dodd, *University of Texas at Austin*
George Engelhard, Jr., *Emory University*
William Fisher, Jr., *Louisiana State University Medical Center*
Carl V. Granger, *SUNY Buffalo*
Gunnar Grimby, *Sahlgrenska University Hospital, Göteburg, Sweden*
Thomas Haladyna, *Arizona State University West*
Allen Heinemann, *Rehabilitation Institute of Chicago*
Robert Hess, *Arizona State University West*
William Koch, *University of Texas at Austin*
Joanne Lenke, *The Psychological Corporation*
J. Michael Linacre, *University of Chicago*
David McArthur, *UCLA School of Public Health*
George Marcoulides, *California State University, Fullerton*
Carol Myford, *Educational Testing Service*
David Nichols, *SPSS, Inc.*
Nambury Raju, *Illinois Institute of Technology*
Randall Schumacker, *University of North Texas*
Mary Segal, *Moss Rehabilitation Research Institute*
Everett Smith, *University of Illinois at Chicago*
Alan Tennant, *University of Leeds, England*
Benjamin D. Wright, *University of Chicago*
David Zurakowski, *Children's Hospital, Boston*

JOURNAL OF OUTCOME MEASUREMENT®, 2(4), 285 - 301

Rasch Measurement for Reducing the Items of the Nottingham Health Profile

Luis Prieto
Health Services Research Unit, Institut Municipal d'Investigació Mèdica (IMIM) and Facultat de Psicologia i Ciencies de l'Educació Blanquerna, Universitat Ramon Llull.

Jordi Alonso, Rosa Lamarca
Health Services Research Unit, Institut Municipal d'Investigació Mèdica (IMIM)

Benjamin D. Wright
The University of Chicago

The present study aimed to develop a short form of the Spanish version of the Nottingham Health Profile (NHP) by means of Rasch analysis. Data from several Spanish studies that included the NHP since 1987 were collected in a common database. Forty-five different studies were included, covering a total of 9,419 subjects both from the general population and with different clinical pathologies. The overall questionnaire (38 items) was simultaneously analyzed using the dichotomous response model. Parameter estimates, model-data fit and separation statistics were computed. The items of the NHP were additionally regrouped into two different scales: Physical (19 items) and Psychological (19 items). Separated Physical and Psychological parameter estimates were produced using the simultaneous item calibrations as anchor values. Misfitting items were deleted, resulting in a 22 item final short form (NHP22) -11 Physical and 11 Psychological-. The evaluation of the item hierarchies confirmed the construct validity of the new questionnaire. To demonstrate the invariance of the NHP22 item calibrations, Rasch analyses were performed separately for each study included in the sample and for several sociodemographic and health status variables. Results confirmed the validity of using the NHP22 item calibrations to measure different groups of people categorized by gender, clinical and health status.

Requests for reprints should be sent to Luis Prieto, Unitat de Recerca en Serveis Sanitaris Institut Municipal d'Investigació Mèdica (IMIM) C/ Dr. Aiguader, 80, Barcelona 8003, Spain. E-mail: lprieto@imim.es

Introduction

Any health problem can place substantial limitations on the normal development of physical, emotional and social aspects of a patient's life. The increasing interest in measures reflecting the personal viewpoint of patients has led to an extended demand for reliable and valid standardized questionnaires of health related quality of life (HRQOL) (Guyatt et al., 1994). Currently, a number of HRQOL instruments are being used to assess the health status of both individuals and populations (McDowell et al., 1987). These questionnaires can differ in numerous manners (I. e., goals, content, methods, culture of origin ...). Considering aspects such as item content, scope and target population, instruments can be basically classified as generic or disease-specific (Guyatt et al., 1991). Each category has its advantages and disadvantages. Specific questionnaires improve the sensitivity of the measurement because they are specially designed to be used for a particular disease (Stucki et al., 1995), but they are not applicable to the general population or any other condition. Generic health measures (e.g. SF-36 Health Survey (Ware et al., 1993), Nottingham Health Profile (European Group for Quality of Life Assessment and Health Measurement, 1993), Sickness Impact Profile (Bergner et al., 1976)) make comparisons possible between different conditions.

The Nottingham Health Profile (NHP) is a generic measure of subjective health status focusing on distress that was originally developed in Great Britain in the late 1970s and which is used extensively in several European countries (European Group for Quality of Life Assessment and Health Measurement, 1993). It contains 38 items with a 'yes/no' response format, describing problems on six health dimensions (Energy, Pain, Emotional Reactions, Sleep, Social Isolation and Physical Mobility) -See appendix A-. The Spanish version of the questionnaire was obtained after an accurate translation process aimed at achieving conceptual equivalence (Alonso et al., 1990). It has proved to be valid and reliable in several groups of patients (Permanyer-Miralda et al., 1991; Alonso et al., 1992; Alonso et al., 1994; Badia et al., 1994). The authors of the original version assigned weights to each NHP item in order to appropriately address the clear disparity in the magnitude of the problems described by each item. Within each dimension (scale), items were weighted using the Case V of the Law of Comparative Judgment (LCJ) proposed by Thurstone (Mckenna et al., 1981). NHP item weights were also replicated for the Swedish (Hunt et al., 1987), French (Bucquet et al., 1990) and Spanish

(Prieto et al., 1996) versions of the questionnaire in order to assess cross-cultural equivalence and validate the adaptation process. However and since the model used for obtaining the Spanish item weights did not fit the data for 5 dimensions of the questionnaire, the use of an unweighted NHP scoring has been recommended for the Spanish version (Prieto et al., 1996). Scores for this purpose are obtained by summing the number of affirmative answers in each scale in the questionnaire and expressing this number as a percentage, range 0 (best health status) to 100 (worst health status).

Although the Spanish version of the NHP has proved to be valid and reliable under a measurement model developed by fiat (Torgerson, 1958), the support for a dimensional perspective of health related problems is weak. Previous work has not evaluated whether the items of the NHP dimensions form a hierarchical item continuum, whether the items represent a single dimension, nor whether the item hierarchy is reproducible across different samples of patients and test occasions. Validation of these scaling properties is a necessary requirement for objective measurement (Wright et al., 1979). The unidimensionality of the NHP component scales has been simply based on the interpretation of Cronbach's Alpha, inter-item correlations and summation of raw scores (Alonso et al., 1994).

In addition to these facts, the length of the NHP (38 items) could be a barrier in using it for clinical purposes since the questionnaire might require excessive patient or physician time. The profile structure of the questionnaire can also bar its use in clinical settings given the difficulty in providing a global interpretation of the 6 different scores offered by the instrument. As a result, the question arises whether it would be possible to develop a shorter version of the NHP, based on a single summary score (index), so that it would be reliable and valid under the perspective of objective measurement.

The Danish mathematician Georg Rasch proposed an alternative scaling approach based in the logistic response model that fully warrants objective measurement (Rasch, 1960). The Rasch analysis builds a variable continuum based on the responses of persons in the sample to the items in the scale, such that persons with "more health impairment" have higher probabilities for giving responses of "limitation" to items than persons with "lower health impairment".

The present study aimed to develop a short form of the Spanish version of the NHP by means of Rasch analysis with a large national database of Spanish patients and non-patients. Selection of items was guided

in such a way that the final questionnaire forms a hierarchical and unidimensional index, reproducible across multiple patient groups. Specifically, we examined: (1) how well the items of the short version contribute to define a single "health" variable, (2) how effective the items are in defining this variable, (3) how the items arrange on the variable continuum, and (4) the invariance of item calibrations across patients with different conditions, levels of perceived health status and gender.

Data and Methods

Subjects

Data collection intended to gather all the studies that had included the Spanish version of the NHP since it was released for general use in 1987. Studies were identified by searching Medline and the Spanish Medical Index from 1987 to 1995 (Key terms: Nottingham Health Profile, NHP, quality of life, measure of health status, questionnaire, reliability, validity, Spanish, and Spain). Studies were also identified by inspecting the Spanish NHP "cession of use" registry maintained since 1987 by one of the authors of the present paper (JA). From the 119 identified NHP studies, data was only available in 45 of them, covering a total of 9,419 individuals. In all studies, the Spanish version of the NHP had been administered. Also, we collected additional information on age, gender, and general health status (response choices: Very good, Good, Fair, Poor, Very poor), among others through a questionnaire to the principal investigator of each study.

Methods

The complete version of the Nottingham Health Profile (NHP38) was consecutively item analyzed following the Rasch dichotomous response model. The dichotomous response model (Wright et al., 1979), suitable for the *Yes/No* response choices of the NHP items, specifies through log-odds that the probability of response of person n to the item i is governed by the location B of the subject (person measure) and the location D of the item (item calibration) along a common measurement continuum:

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

where, P_{ni1} is the probability of a *Yes* response to the item i and P_{ni0} is the probability of a *No* response. When $B_n>D_i$, there is more than a 50% chance

of a *Yes* response. When $B_n=D_i$, the chance for a *Yes* response is 50%. When $B_n<D_i$, the probability is less than 50%.

Each facet in the model (B,D) is a separate parameter. The effects of one parameter are free from the effects of the others (Rasch, 1960; Wright et al., 1979; Wright et al., 1982). This mathematical property enables "test-free" and "person-free" measurement to occur, a prerequisite for objective measurement. "Test-free" means that person measures do not depend on which items are used to measure them. "Person-free" means that item estimates do not depend on which sample is being measured.

The item calibrations define the hierarchical order of severity of the NHP items along the health continuum. Item calibrations are expressed in log-odd units (logits) that are positioned along the hierarchical scale. A logit is defined as the natural log of an odds ratio. An odds-ratio for an item is the level of severity of the item in relation to the severity of the total set of items, with logits of greater magnitude representing increasing item severity. The unidimensionality of the scale is determined by the pattern of item goodness-of-fit statistics. The goodness-of-fit statistics compare each person's observed responses to the expected response pattern for each specific overall score (Wright et al., 1979).

Rasch analyses were performed using the computer program BIGSTEPS version 2.73 (Wright et al., 1997). To avoid negative values, BIGSTEPS estimates were rescaled in Response Probability Scaling Units (Chips) in all the analyses by establishing a new origin (50 units) and spacing (9.1 units/1 logit) of the scale (Wright et al., 1979). In order to determinate the spacing of each estimate, an associated standard error (SE) was calculated for each item. Other separation indices, indicating the extent to which items and persons identify a useful variable line, were also calculated (Wright et al., 1982). The person separation index gives the sample standard deviation in standard error units: it equals the square root of the ratio of true variance of person measures to the error variance due to person measurement imprecision. The item separation index indicates how well items spread along the variable line by giving the item standard deviation in calibration error units. Reliability (R) of person and item separation is provided by the relationship between R and separation (SEP):

$$R = (SEP)^2 / (1+SEP)^2$$

Infit and outfit mean square statistics (MNSQ) were used to determine how well each NHP item contributed to define the common health

variable (Goodness-of-fit test) (Wright et al., 1979). Infit identifies unexpected responses of items close to the respondent's measure levels. Outfit detects unexpected responses to items which are distant from the respondent's measure levels. An item with a MNSQ near 0 indicates that the sample is responding to it in an overly predictable way. Item MNSQ values of about 1 are ideal by Rasch model specifications, since it indicates local independence. Items with MNSQ values greater or equal to 1.3 were diagnosed as potential misfits to Rasch model conditions and deleted from the assessed sequence. Successive Rasch analyses were performed until a final set of items satisfied the model fit requirements.

A single summary score of the NHP has the advantage of simple interpretation at the expense of ability to detect different patterns of health impairment. In order to lessen the potential loss of sensitivity of the new short questionnaire, two additional scoring options were taken into account. Considering results of principal components analysis of residuals from expectation as well as previous experience with the questionnaire, the 38 items of the NHP were regrouped into two new scales before further Rasch analyses were performed: a Physical Scale (containing Energy, Pain and Physical Mobility dimensions) -19 items- and a Psychological Scale (containing Emotional Reactions, Sleep and Social Isolation) -19 items-. Physical and Psychological items were jointly calibrated in order to investigate improvements in scale definition when all items were in the same unit of measurement. Parallel to each consecutive Rasch analysis of the items, separate Physical and Psychological item and person measures were also produced using the simultaneous item calibrations as anchor values. The displacement of each estimate away from the statistically better value, which would result from the best fit of the data to the model, was provided for each Physical and Psychological item. Pearson's correlation coefficients were calculated between NHP22, Physical and Psychological person measures.

To determine whether the final set of item calibrations was invariant, Rasch analyses were performed separately for each substudy included in the common database, as well as by gender, age and health severity groups. For this purpose, two severity groups expected to differ in physical and mental health status were defined. Using information provided by the studies under analysis, two mutually exclusive disease-severity groups were formed: Patients and Non-patients. Additionally, when the information was available in the original study (50% of the overall sample), individuals were also classified in another two health severity groups: those

reporting a "Very good" or "Good" general health status (Group 1), and those indicating "Fair", "Poor" or "Very poor" health (Group 2). In order to perform an additional validation study on the stability of the item calibrations of the new short form, subjects in the initial common database were randomly split into two different subsamples. Main analyses described above were performed in subsample A (85%, n=8,015)) and repeated in subsample B (15%, n=1,404).

Results and Discussion

Table 1 presents the main characteristics of the population in the common database created from the 45 studies. The mean age of the overall sample was 57 (range 12 to 99). Although significant differences were found for gender distribution (probably due to the high sample size), nearly 50% of the sample were female. Subjects came from the general population or were patients with different clinical pathologies. Around 50% of the dataset were made up of non-patients.

The simultaneous Rasch analyses of the 38 items of the NHP showed 9 misfitting items (PM8, EN1, EN2, P7, SL1, SL2, SO3, EM8, and EM5): INFIT MNSQ statistics ranged from 0.78 to 1.30 (standard deviation=0.14) and OUTFIT MNSQ ranged from 0.62 to 2.39 (standard deviation=0.41). These 9 as well as other misfitting items detected over the course of 5 subsequent Rasch analyses were successively removed until there was no further improvement in the fit requirements. In this process, 16 different items were erased from the analyses, finally reducing the initial questionnaire to 22 items (NHP22). The component item calibrations, standard errors and fit statistics of NHP22 are reported in Table 2. A map of items and persons distribution is provided in Figure 1. There were 6,052 people measurable for this Rasch analysis. Data was missing for 456 persons. Also, 1,956 individuals were deleted from the overall analysis since they either reported no health problem (n=1,361), or indicated every health problem contained in the questionnaire (n=146) . Items are arranged in more severe to less severe health status. The items varied in severity from 31.11 to 70.28 chips with standard errors of .28 to .49. Eighteen of the 22 items fit to define a unidimensional variable according to Rasch specifications (INFIT and OUTFIT MNSQ < 1.3). The 4 items which misfit (OUTFIT MNSQ 1.3) were retained for practical considerations discussed below. The standard deviation of the INFIT and OUTFIT MNSQ dropped to 0.09 and 0.24 respectively. Person separation for the NHP22 was 2.08 (R=0.81) and item separation was 27.51 (R=1).

Table 1
Characteristics of the population used in the study

	ALL n=9,419	MALES† n=4,478	FEMALES† n-4,908	p-VALUE‡ males Vs females
GENDER (%)		47.5	52.1	<0.001
AGE GROUPS (%)				
12-44	24.5	23.1	26.0	0.001
45-54	14.9	15.5	14.5	0.179
55-64	18.9	21.6	16.5	<0.001
65-74	24.6	25.6	23.7	0.033
75-99	16.7	14.1	19.2	<0.001
STUDY POPULATIONS (ICD-9 CODES) (%)				
Non-patients*	49.8	44.1	55.3	<0.001
Diseases of the musculoskeletal system & connecttive tissue	13.2	7.1	18.6	<0.001
Respiratory system	10.2	15.6	5.0	<0.001
Genitourinary system	7.6	9.4	6.0	<0.001
Procedures	6.3	6.8	5.8	0.045
Nervous system & sense organs patients	5.6	5.6	5.6	0.997
Circulatory system	1.8	3.3	0.4	<0.001
Mental disorders	1.4	2.1	0.9	<0.001
Digestive system	1.4	1.6	1.2	0.094
Neoplasm	1.0	1.9	0.1	<0.001
Endocrine, nutritional & metabolic diseases & immunity disorders	0.8	0.9	0.7	0.273
Infectious and parasitic diseases	0.8	1.3	0.5	<0.001

† A subset of individuals (n=33) did not report information of their gender.
‡ Based on chi-square test
* This group included general population (different age groups) and pregnant females.

The item hierarchies of the Physical and Psychological subscales of the NHP22 are reported in Table 3. Maps of items and persons distribution appear of Figure 2. The displacement of the estimates from the anchored measure of the simultaneous calibration was slight (Range, in absolute values, 0 to 3.8 chips for the Physical Scale and 0 to 1.3 chips for the Psychological Scale). Nine of the 11 Physical items and 10 of the 11 Psychological items fit to define a unidimensional variable. Person separation was 1.39 (R=0.66) and item separation was 30.23 (R=1) for the Physical scale . For the Psychological scale, person separation was 1.24 (R=0.61) and item separation was 22.05 (R=1). The 3 items which misfit (PM1and PM4 for Physical, and EM1 for Psychological) were the same which misfitted in the simultaneous calibration of Table 2. The outfit statistics indicate there were a few unexpectedly high and low scores across individuals for these items. Considering (1) that their extreme positions

Table 2
NHP22 item hierarchy (n=6,052 persons)

NHP22 ITEMS	MEASURE	SE	INFIT MNSQ	OUTFIT MNSQ
PM3-UNABLE WALK	SICK 70.28	.49	.93	1.03
SO5-PEOPLE HARD	64.52	.43	1.08	1.40
PM1-WALK LIMITED	61.04	.39	1.07	1.30
SO2-CONTACT HARD	58.37	.37	1.06	1.28
P2 -AWFUL PAIN	57.65	.35	.91	.86
SO4-IM A BURDEN	56.81	.34	.97	.95
EM4-DAYS DRAG	54.87	.33	.92	1.02
PM6-HARD TO DRESS	54.60	.33	.91	.77
EM6-NO CONTROL	54.01	.33	.96	.96
P8 -SITTING PAIN	52.59	.32	.96	.89
EM9-DEPRESSED	48.76	.30	.95	.92
PM5-REACH HARD	48.25	.30	.88	.76
EM2-JOY FORGOTTEN	46.31	.30	1.05	1.13
EN3-OUT OF ENERGY	46.16	.30	.88	.81
SL3-CANT SLEEP	45.57	.29	1.00	.94
P3 -CHANGE PAIN	43.65	.29	.95	.99
SL5-SLEEP BADLY	43.14	.29	.96	.96
SL4-SLOW TO SLEEP	41.93	.30	1.11	1.26
P4 -WALK PAIN	41.81	.29	.98	1.00
EM1-GETTING DOWN	40.02	.32	1.24	1.55
PM2-HARD TO BEND	38.54	.28	.99	1.03
PM4-STAIRS HARD	WELL 31.11	.30	1.06	1.68
MEAN	50.00	.33	.99	1.07
SD	9.16	.05	.09	.24

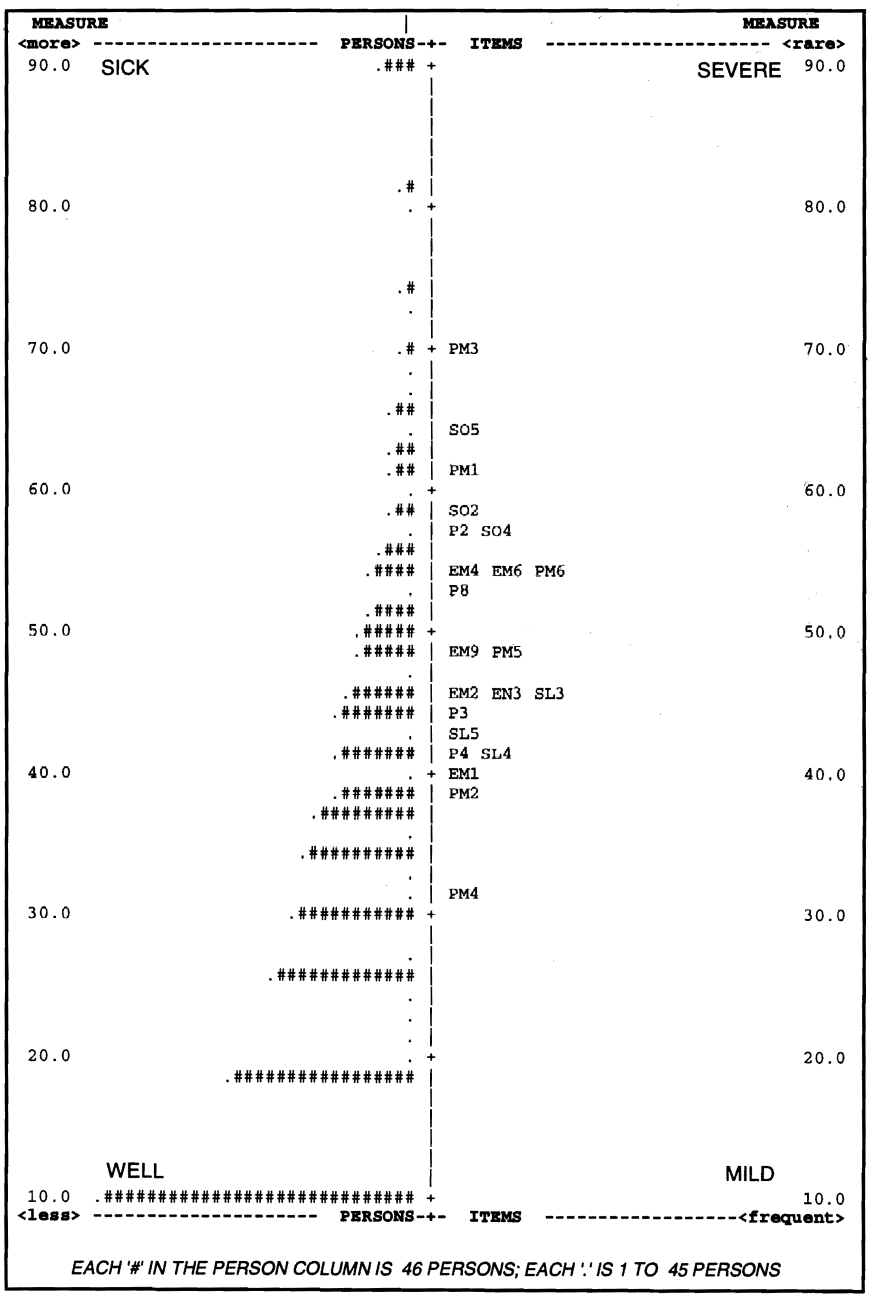


Figure 1. Map of persons (n=6,052) and NHP22 items

in the hierarchy are nevertheless conceptually valid and (2) that their exclusion substantially decreased the person separation index of the questionnaire these misfitting items were retained here and in the simultaneous analysis in Table 2.

The person measures from the separate calibration of Physical items correlated 0.92 with the simultaneous calibrated person measures. For the person estimates from the Psychological items this correlation was 0.91. Physical and Psychological measures correlated 0.70. These correlations demonstrate that the Physical and Psychological person estimates are substantially associated with each other and highly associated with the simultaneously calibrated person estimates, indicating a common underlying construct.

From Tables 2 and 3, a logical sequencing of the health problems depicted by the items can be observed (Figures 1 and 2), supporting the construct validity of the obtained responses. Mild health problems such

Table 3
NHP22 Physical and Psychological item hierarchies

PHYSICAL SCALE (n=5,003 persons)						
ITEMS	ANCHORED MEASURE		SE	DISPLACE	INFIT MNSQ	OUTFIT MNSQ
PM3-UNABLE WALK	SICK	70.3	.5	3.2	.82	.96
PM1-WALK LIMITED		61.0	.4	1.8	1.03	1.35
P2 –AWFUL PAIN		57.7	.4	1.5	.89	.92
PM6-HARD TO DRESS		54.6	.3	1.2	.83	.72
P8 –SITTING PAIN		52.6	.3	.9	.93	.90
PM5-REACH HARD		48.3	.3	.3	.83	.74
EN3-OUT OF ENERGY		46.2	.3	0.0	.99	.99
P3 –CHANGE PAIN		43.7	.3	-.4	.96	.97
P4 –WALK PAIN		41.8	.3	-.8	.92	.91
PM2-HARD TO BEND		38.5	.3	-1.5	.90	.93
PM4-STAIRS HARD	WELL	31.1	.3	-3.8	.99	1.50
PSYCHOLOGICAL SCALE (n=4,984 persons)						
ITEMS	ANCHORED MEASURE		SE	DISPLACE	INFIT MNSQ	OUTFIT MNSQ
SO5-PEOPLE HARD	SICK	64.5	.4	1.3	1.00	1.17
SO2-CONTACT HARD		58.4	.4	.6	.96	1.03
SO4-IM A BURDEN		56.8	.4	.6	.98	1.02
EM4-DAYS DRAG		54.9	.3	.5	.90	.89
EM6-NO CONTROL		54.0	.3	.4	.90	.88
EM9-DEPRESSED		48.8	.3	0.0	.86	.81
EM2-JOY FORGOTTEN		46.3	.3	-.2	1.06	1.11
SL3-CANT SLEEP		45.6	.3	-.2	.93	.88
SL5-SLEEP BADLY		43.1	.3	-.5	.89	.86
SL4-SLOW TO SLEEP		41.9	.3	-.6	1.02	1.08
EM1-GETTING DOWN	WELL	40.0	.3	-.8	1.20	1.34

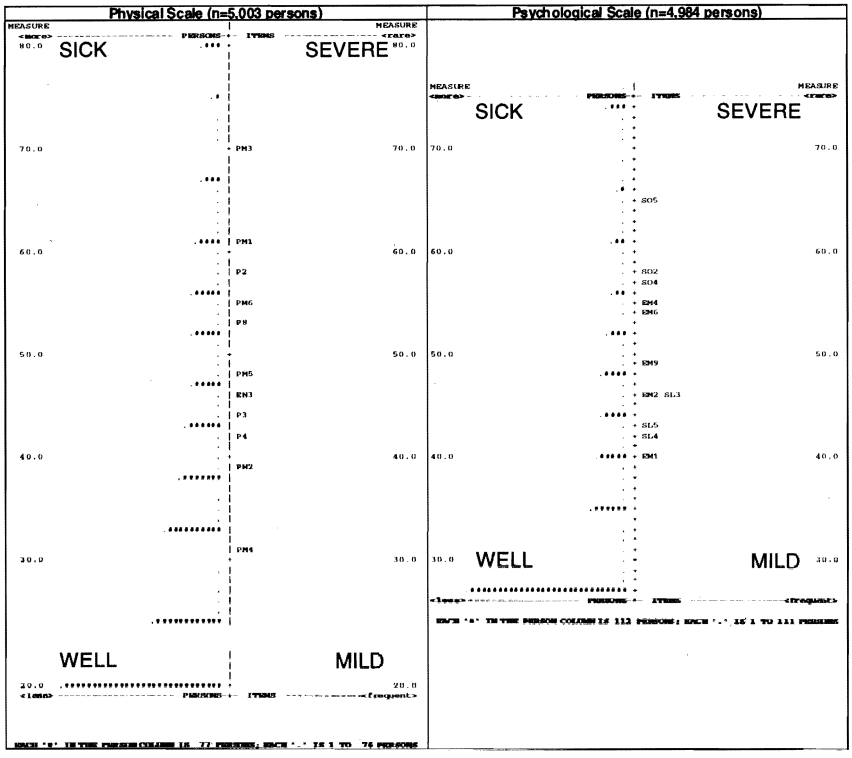


Figure 2. Map of persons and NHP22 items by scale

as PM4 "I have trouble getting up and down stairs" and EM1 "Things are getting me down" are easy to endorse and are found in the bottom of the health severity continuum. More severe items such as PM3 "I'm unable to walk at all" and SO5 "I'm finding it hard to get on with people", are found in the top of the continuum. The substantial and reliable person separation (PS) indices obtained for each measure (PS range 1.24 to 2.08; reliability range 0.61 to 0.81) suggest that the sample of individuals in the study was well targeted by the questionnaire. Person separation indexes produced 3 statistically distinct person strata (Wright et al., 1982) identified by the NHP22 scale, and 2 for the Physical and the Psychological scales. Item separation indexes (range 22.05 to 30.23; reliability = 1) also indicate a good and reliable separation of the items along the variables which they define.

The multiple box plot in Figure 3 summarizes information about the distribution of the NHP22 item calibrations across the different studies composing the sample. All the item calibrations for each study were stan-

dardized by subtracting the value of the corresponding item measure in Table 2, and thus expressed in units of displacement from these values. The lower boundary of each box is the 25th percentile, and the upper boundary is the 75th percentile. The horizontal line in the box represents the median. The box plot includes two categories of cases with outlying values. Extreme values (crosses) are cases with values more than 3 box-lengths from the upper or lower edge of the box. Outliers (squares) are cases with values between 1.5 and 3 box-lengths from the edge of the box. Lines are drawn from the ends of the box to the largest and smallest observed values that are not outliers. The majority of the items showed median displacements close to 0 as well as similar distributions of calibrations. The spread of non-outliers displacements was slight, mainly falling between ± 10 units of displacement (± 1.1 logits), suggesting the invariance of item calibrations across different groups of individuals. The larger spread of item scores was observed between the items located in the extremes of the continuum, especially for those items showing misfit in Table 2 (SO5, PM1, EM1 and PM4). Extreme values and outliers varied between ± 30 units of displacement (± 3.3 logits). Only one item (PM3)

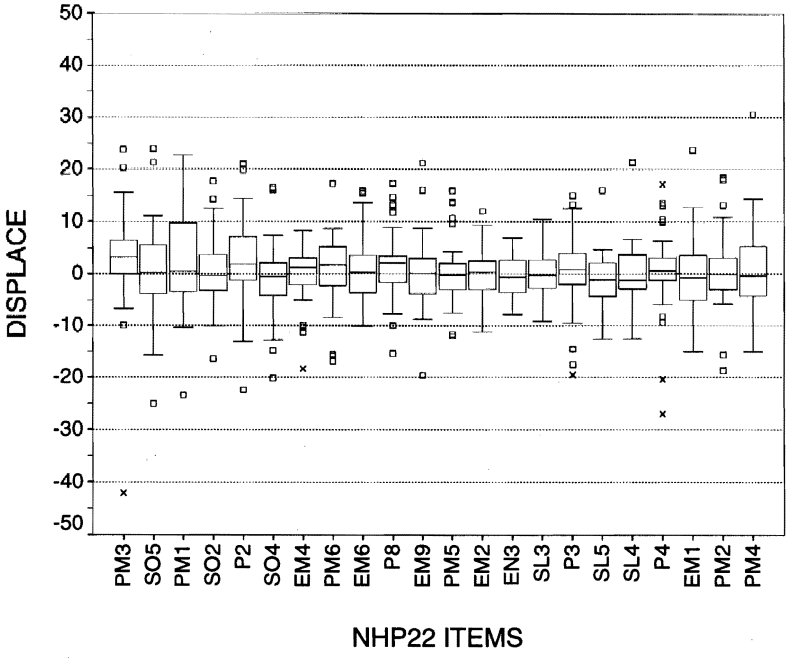


Figure 3. NHP22 item scores by study

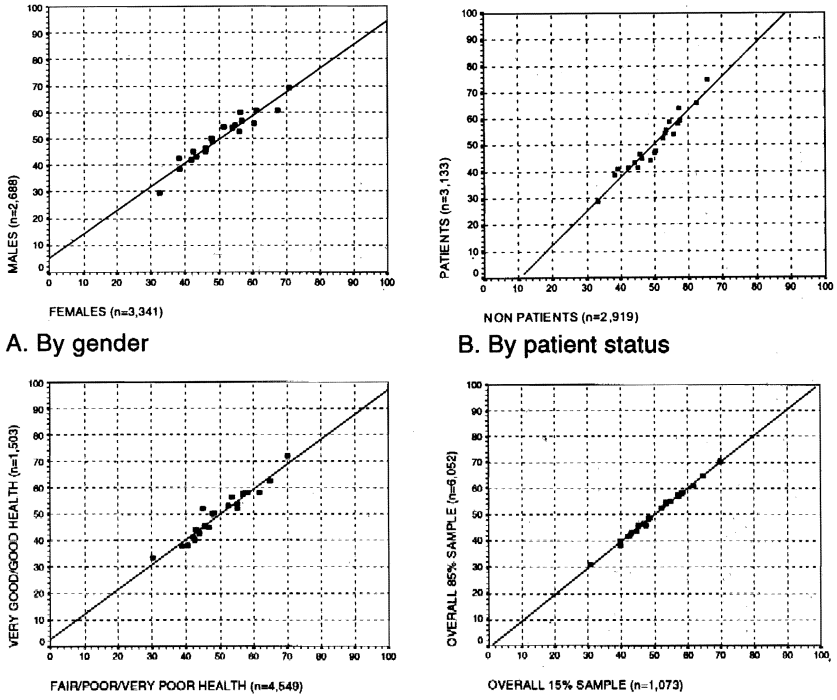


Figure 4. Correlation of NYP22 item calibrations, according to: A. Gender, B. Patient status, C. General health status, D. Sample

showed an extreme displacement beyond 40 units. Most of the extreme values and outliers were due to a minor group of studies, basically characterized for their composition of young individuals without serious health problems, for whom the questionnaire would be too severe.

The item plots in Figure 4 compare the severity of NHP22 items obtained by gender, patient status, general health and sample of analysis. Pearson's correlation coefficients were 0.96 for gender, 0.97 for patient status, 0.96 for general health and 1.00 for sample, placing the NHP22 item severities between categories close enough to an identity line to conclude that they are comparable across the defined categories of each group. Differences in item calibration between samples (85% Vs 15%) were almost nonexistent, confirming the validity of the overall item calibrations in Table 2.

Conclusion

The Spanish version of the Nottingham Health Profile (NHP) was shortened to 22 items (NHP22) by means of Rasch analysis. Through goodness-

of-fit statistics and the investigation of the hierarchy of item calibrations, a unidimensional view of health of the NHP22 was confirmed. Physical and Psychological measures were also calculated for the NHP22. In this way, the NHP22 offers the possibility to be scored in three modes: (1) a global score based on the 22 items, (2) a Physical score based on only 11 items and (3) a Psychological score based on the remaining 11 items. Given the adequate measurement properties shown by the Physical and Psychological items, both scales might be independently administered to individuals. These alternatives in scoring provide clear advantages in understanding and presenting results over the original six-dimension profile of the NHP.

Although there were some differences in the hierarchical structures of the various sets of NHP22 item calibrations obtained by study, gender, patient and health status, the significance of these discrepancies was slight. Thus, our results could be easily generalized to males and females with different health status, ranging from general population to those clinical pathologies included in the study. The item measures in Table 2 should be used to measure new people.

Further research is nevertheless necessary to confirm the utility and stability of the item calibrations found for the NHP22. An important limitation in our study is that all the data used to select the items of the NHP22 were based on information gathered using the original NHP of 38 items. Although unlikely, it is possible that the framework of the total NHP might have influenced the results found. Using the NHP22 as an independent instrument will have to reproduce results as well as to prove its utility as a generic measure of health related quality of life. However, our conclusion is that the 22-item selection provides a promising short alternative for the original NHP under the perspective of objective measurement.

Acknowledgement

The authors gratefully acknowledge Dave MacFarlane for his editorial assistance.

References

Alonso, J., Antó, J. M., and Moreno, C. (1990). Spanish Versión of the Nottingham Health Profile: Translation and Preliminary Validity. *American Journal of Public Health*, *80*, 704-708.

Alonso, J., Antó, J. M., González, M., Fiz, J. A., Izquierdo, J., and Morera, J. (1992). Measurement of general health status of non-oxygen-dependent chronic obstructive pulmonary disease patients. *Medical Care*, *30*, MS125-MS135.

Alonso, J., Prieto, L., and Antó, J. M. (1994). The Spanish version of the Nottingham Health Profile: a review of adaptation and instrument characteristics. *Quality of Life Research*, *3*, 385-393.

Badia, X., Alonso, J., Brosa, M., and Lock, P. (1994). Reliability of the Spanish Version of the Nottingham Health Profile in Patients with Stable End-Stage Renal Disease. *Social Science and Medicine, 38*, 153-158.

Bergner, M., Bobbit, R. A., and Pollard, W. E. (1976). The sickness impact profile: validation of a health status measure. *Medical Care, 14*, 57-67.

Bucquet, D., Condon, S., and Ritchie, K. (1990). The French Version of the Nottingham Health Profile. A Comparison of items Weights with Those of the Source Version. *Social Science and Medicine, 30*, 829-835.

European Group for Quality of Life Assessment and Health Measurement. (1993). *European guide to the Nottingham Health Profile*. Surrey: Brookwood Medical Publications.

Guyatt, G., Feeny, D., and Patrick, D. (1991). Issues in quality-of-life measurement in clinical trials. *Controlled Clinical Trials, 12*, 81S-90S.

Guyatt, G. H., and Cook, D. J. (1994). Health status, quality of life, and the individual. *Journal of the American Medical Association, 272*, 630-631.

Hunt, S. M., and Wiklund, I. (1987). Cross-cultural variation in the weighting of health statements: a comparison of English and Swedish valuations. *Health Policy, 8*, 227-235.

McDowell, I., and Newell, C. (1987). *Measuring health: A guide to rating scales and questionnaires*. New York: Oxford University Press.

Mckenna, S. P., Hunt, S. M., and McEwen, J. (1981). Weighting the seriousness of perceived health using Thurstone's method of paired comparisons. *International Journal of Epidemiology, 10*, 93-97.

Permanyer-Miralda, G., Alonso, J., Antó, J. M., Alijarde-Guimerá, M., and Soler-Soler, J. (1991). Comparison of perceived health status and conventional functional evaluation in stable patients with coronary artery disease. *Journal of Clinical Epidemiology, 44*, 779-786.

Prieto, L., Alonso, J., Viladrich, M. C., and Antó, J. M. (1996). Scaling the Spanish version of the Nottingham Health Profile: evidence of limited value of item weights. *Journal of Clinical Epidemiology, 49*, 31-38.

Rasch, G. (1960, 1993). *Probabilistic models for some intelligence and attainment tests.* Chicago: MESA Press. Original Edition, the Danish Institute for Educational Research.

Stucki, G., Feeny, D., and Patrick, D. (1995). Issues in quality-of-life measurement in clinical trials. *Controlled Clinical Trials, 48*, 1369-1378.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

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

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

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

Wright, B. D., and Linacre, J. M. (1997). *A user's guide to BIGSTEPS: Rasch-Model Computer Program, version 2.7*. Chicago: MESA Press.

Appendix A

THE 38 ITEMS OF THE NOTTINGHAM HEALTH PROFILE

ENERGY
EN1 I'm tired all the time
EN2 Everything is an effort
EN3 I soon run out of energy

PAIN
P1 I have pain at night
P2 I have unbearable pain
P3 I find it painful to change position
P4 I'm in pain when I walk
P5 I'm in pain when I'm standing
P6 I'm in constant pain
P7 I'm in pain when going up or down stairs
P8 I'm in pain when I'm sitting

EMOTIONAL REACTIONS
EM1 Things are getting me down
EM2 I've forgotten what it's like to enjoy myself
EM3 I'm feeling on edge
EM4 These days seem to drag
EM5 I lose my temper easily these days
EM6 I feel as if I'm losing control
EM7 Worry is keeping me awake at night
EM8 I feel that life is not worth living
EM9 I wake up feeling depressed

SLEEP
SL1 I take tablets to help me sleep
SL2 I'm waking in the early hours of the morning
SL3 I lie awake for most of the night
SL4 It takes me a long time to get to sleep
SL5 I sleep badly at night

SOCIAL ISOLATION
SO1 I feel lonely
SO2 I'm finding it hard to make contact with people
SO3 I feel there is nobody I am close to
SO4 I feel I am a burden to people
SO5 I'm finding it hard to get on with people

PHYSICAL MOBILITY
PM1 I can only walk about indoors
PM2 I find it hard to bend
PM3 I'm unable to walk at all
PM4 I have trouble getting up and down stairs
PM5 I find it hard to reach for things
PM6 I find it hard to dress myself
PM7 I find it hard to stand for long
PM8 I need help to walk about outside

JOURNAL OF OUTCOME MEASUREMENT®, 2(4), 302 - 325

Generalizability Theory: A Unified Approach to Assessing the Dependability (Reliability) of Measurements in the Health Sciences

Dawn M. VanLeeuwen
Agricultural Biometric Service
New Mexico State University

Michael D. Barnes
Department of Health Sciences
Brigham Young University, Provo

Marilyn Pase
Nursing Department
New Mexico State University, Las Cruces

The reliability of health promotion program evaluation measures, behavioral and attitudinal measures, and clinical measures is a concern to many health educators. Classical reliability coefficients, such as Cronbach's alpha, apply to narrowly defined, prespecified measurement situations. Classical theory does not provide adequate reliability assessments for criterion-referenced measures, for measurement situations having multiple sources of error, or for aggregate-level variables. Generalizability theory can be used to assess the reliability of measures in these situations that are not adequately modeled by Classical theory. Additionally, Generalizability theory affords a broader view and a deeper understanding of the dependability of measurements and the role of different sources of error in the variability of measures.

Requests for reprints should be sent to Dawn VanLeeuwen, Agricultural Biometric Service, Box 30003, MSC 3130, New Mexico State University, Las Cruces, NM 88003-8003

Introduction

Health educators and health education researchers obtain or interpret several types of measurements including measurements used to evaluate health promotion programs, behavioral and attitudinal measurements, and clinical measurements. For these types of measurements, the reliability or dependability of the measurement is a concern in the measurement's utility and interpretation. Generally the need to assess a measurement's reliability is recognized and, typically, coefficients from Classical Theory (CT) such as Cronbach's alpha, the test-retest coefficient of stability, the Spearman-Brown prophecy formula, or Kuder-Richardson formula 20, are used. These coefficients are often applied in measurement situations where they are either inappropriate or inadequate. For example, CT is inappropriate when the measure is criterion-referenced (Schaeffer et al., 1986), when an aggregate-level variable is the variable of interest (O'Brien, 1990), or in the presence of multiple sources of error (Eason, 1989). As an alternative, Generalizability Theory (GT), can provide a unified framework for examining the dependability of measurements and is appropriate whether the measurement is norm- or criterion-referenced, whether the measurement is an individual-level or an aggregate-level variable, and in the presence of single or multiple sources of error.

Superficially, it would appear that the strongest advantage GT offers over CT is that it can model and estimate the variability due to a number of different sources of error, including interactions, simultaneously. Estimates of the various sources of error can then either be interpreted or be used in determining a reliability coefficient. But GT's value as a unified framework within which to examine the reliability of measurements goes beyond this. GT's unified framework for examining the reliability of measurements for both relative and absolute decisions has a particular value to health practitioners because criterion-referenced measures are prevalent in the health sciences (Schaeffer et al., 1986). In addition, GT applies the same methodology to all sources of error and all possible objects of measurement. Thus, the same framework can be used to examine the dependability of measurements whether the object of measurement is organizations or individuals, and whether items, test forms, occasions, raters, or any combination of these are the sources of error. Thus, unlike CT with its multitude of formulas, GT represents a unified approach to examining the dependability of measurement processes.

GT's unified approach relies on the powerful and flexible machinery of random and mixed linear models (Kempthorne, 1957; Scheffè, 1959; Searle, 1971) and variance components estimation. GT emphasizes estimating and interpreting variance components rather than simply obtaining a single number index of reliability. This emphasis on the estimation of variance components is rooted in GT's focus on understanding the variability inherent in measurement processes or, in GT terms, assessing the "dependability of measurements". While many applications of linear models focus on applying fixed or mixed linear models to comparing treatment effects, GT emphasizes applying random and mixed linear models to assess measurement error, and to understand the impact of sources of error on the reliability of measurements.

While GT provides a broad view of the dependability of measurements, it is important to recognize that coefficients derived from GT are reliability coefficients (O'Brien, 1995) and there may be disadvantages to using terminology that separates GT from notions of reliability. The phrase "dependability of measurements" is typically associated with the broader notion (Cronbach et al., 1972). However, O'Brien emphasizes the value of appropriately labeling coefficients derived using this theory as "reliability" coefficients.

Within the context of a broader notion of the dependability of measurements, GT places an emphasis on asking the right questions and on appropriately modeling measurements (Kane, 1993). Appropriate use and interpretation of measurements requires simultaneous consideration of the ultimate use of measurements, the quantities or behaviors (true scores) estimated by the measurement, and the manner in which samples are obtained and used to estimate true scores. GT emphasizes the impact of these aspects of the contextual framework on modeling observations and on estimated scores.

The GT literature suggests that this emphasis on context is unique to GT. However, similar emphasis has arisen in any subject area to which linear models or other statistical technical machinery have been extensively applied. For example, Cox (1958) discusses concerns in the design of experiments, while Croxton and Cowden (1955) discuss many issues that are of concern in the collection and interpretation of data. Panse and Sukhatme (1954) also discuss similar issues in planning and interpreting agricultural experiments.

Brennan (1997) notes that (p 17) "Most published generalizability analyses are in the educational literature ..." and that (p 18) "G theory

seems very much underutilized in psychological and medical areas." Even so, GT has been used even less in the health education literature than in the medical and nursing literature where it has been used by some researchers to assess the dependability of various clinical measurements. Some of these measurements such as blood pressure (BP) (Llabre et al., 1988; Szalai et al., 1993) are of interest to health educators, and the interpretation and dissemination of information from these studies lie within the domain of the health educator. Other applications of GT to clinical measurements have involved isometric force measurements (Roebroeck et al., 1993), passive ankle dorsiflexion measurements in children (Watkins et al., 1995), surface electromyographic measurements (Hatch et al., 1992), aortic blood flow measurements using Doppler echocardiography (Kusumoto et al., 1995), and functional performance measurements of Alzheimer patients (Carswell et al., 1995). Other applications of GT in the literature include assessing reliability for patient classification systems (McDaniel, 1994), and proposing guidelines for interpreting indicators of residency program performance (Norcini & Day, 1995). While these applications appear in the medical and nursing literature, health educators deal with similar problems and can apply these same methods to the problems they encounter.

The purposes of this paper are to explicitly consider the different contexts within which GT is applied, and to focus on the power and flexibility of GT's linear models methodology to appropriately model measurement situations that have relevance to the health educator. ANOVA and the underlying linear models have a well-developed methodological machinery available that is not being fully exploited -- even by those who are currently applying these methods to the reliability of measurements. It is our goal to indicate how this machinery might be used by health educators.

We present a brief overview of elementary GT concepts in the next section. Fundamentals of GT are discussed in more detail in the references cited. Shavelson and Webb (1991), Webb et al. (1988), and VanLeeuwen (1997) provide introductions to GT with presentations of numerical examples. Additional introductory papers include Shavelson et al. (1989), Eason (1989), and Brennan and Johnson (1995). Brennan (1983) presents detailed information on the GT analysis of many measurement designs. While Cronbach et al.'s (1972) treatment is comprehensive, their discussion is too technical to be widely accessible to

practitioners (Thompson, 1991) but Shavelson and Webb (1991) provide a treatment that is accessible to practitioners (see reviews by Thompson, 1991; Brown, 1992; Kane, 1993). These references provide numerical examples and technical details on variance components estimation.

GT Basics

GT considers two types of studies: generalizability (G) studies are associated with the development of a measurement process while decision (D) studies obtain measurements for a particular purpose. Information from G studies are used to design D studies that yield measurements having the desired level of reliability for some decision-making purpose. Typically, in GT the object of measurement is people or subjects, although in many applications in health science it may be organization. Kane (1993) notes that a strength of GT is the range of possibilities in defining the object of measurement. Because organizations rather than individuals may be treated as the object of measurement, GT has relevance in a number of health science applications.

Facets are potential sources of error. For example, if BP readings are taken on several days, then averaged to obtain a measurement of a person's underlying BP, the person is the object of measurement, while days are a facet or source of error. If the average is taken using five readings from five different days, then the study includes five conditions (i.e., different days) of the facet days. Similarly, facets may be test items, testing occasions, and, in situations where organizations are the object of measurement, even individuals with different individuals representing different conditions of the facet individual.

Central to GT is the notion of a universe of generalization. Specification of the universe of generalization is the responsibility of the decision-maker (Brennan, 1983). As noted by Kane (1993) this aspect of GT's framework places emphasis on asking the right questions. This framework is different from that of CT and Kane indicates that this emphasis on asking the right questions is one of the major contributions of GT. The universe of generalization is the entire universe of possible observations to which the decision-maker wishes to generalize. Brennan (1983) notes that, based on this universe, the decision-maker must decide which facets are fixed and which are random. Random facets are simply those facets for which universe conditions have been sampled. This sample is either a random sample, or it may be an "exchangeable" sample of conditions. In

some instances, a facet is fixed. This is the case in Llabre et al. (1988) where they include measuring device as a facet. The three conditions of measuring device, ambulatory monitor, mercury sphygmomanometer, and Dinamap, are the only three conditions of measurement device present in the universe of generalization. The technical details of modeling fixed effects are not discussed here, however, Brennan (1983) serves as a reference for the analysis of both random and mixed models. Since the universe score is the object of measurement's average score over all combinations of conditions in the universe of generalization and a test score is based only on a random sampling of the conditions of either some or all of the facets in the universe of generalization, the test score is only an estimate of the universe score.

In many instances, a G study deals with a slightly broader universe than the D study's universe of generalization. This broader universe is called a universe of admissible observations. A single G study, with appropriate sampling from the universe of admissible observations, may provide information that can serve as a basis for planning a number of different D studies each having a different universe of generalization.

GT as an Extension of CT

Kane (1993) notes two ways to think about GT. He states (p 271-2) "One way to think about generalizability theory is as an extension of classical reliability theory with G theory addressing basically the same issues as the classical theory but doing so in a more general and flexible way... Another way to think about G theory is as a framework for thinking about the dependability of measurements, one that encourages the decision maker to formulate the questions that need to be addressed for a specific D study and then to seek answers to those questions." In this section we consider GT as an extension of CT. We introduce the basics of linear modeling and show how these models allow consideration of both relative and absolute decisions as well as allow a multi-faceted model of measurement error. In the next section we consider GT as a framework for thinking about the dependability of measurements.

The single facet pxi design

GT's generality and flexibility is a result of the linear models technical apparatus used to model, think about, interpret, and estimate the variability due to various sources of error. Consider a simple behavioral

score based on using multiple items where the same set of items is administered to each person (i.e., item is crossed with person, denoted pxi). Suppose that a typical CT measure of internal consistency such as Cronbach's alpha, is used to estimate the reliability of the score. The basic CT model states that a given person's response on each question is simply the sum of the person's underlying behavioral characteristic and an independent error. Thus the CT model is

person item score = person's true value + error.

But the GT model for this situation says

person item score = grand mean + person main effect + item main effect + (person x item interaction effect, and error).

While GT's (grand mean + person main effect) is equivalent to CT's (person's true value), CT omits consideration of the item main effect. Thus, this simple situation having only a single facet - items - illustrates a fundamental difference between the GT model and the CT model. Additionally, because of this difference, CT applies only to relative decisions while GT can deal with both relative and absolute decisions.

Relative or norm-referenced assessments indicate one's standing in a score distribution as compared to that of others within the norm group (Isaac and Michael, 1990). That is, relative decisions are essentially based on rankings within some group. GT allows items to vary in difficulty while CT assumes that items are equally difficult or that each item measures the same amount of the underlying attribute for each person. CT ignores the item main effect and considers only error, which in this simple example is not separable from the person x item interaction effect. This error is reflected in changes of rankings of persons when ranked using different items and is not associated with overall differences in item difficulty as is the item main effect term. Thus CT considers only that portion of the error affecting rankings.

Absolute or criterion-referenced assessment reveals one's mastery level of knowledge, attitudes or behavior as compared to specific knowledge, attitude or behavioral criterion (Isaac and Michael, 1990). Schaeffer et al. (1986) note that criterion-referenced measures are common in health-education applications, that CT is not appropriate for such measures, and suggest alternatives, including GT, to assessing reliability of such measures. Item main effect variability has an impact on the absolute score (but not the rankings) and must be considered to correctly assess the reliability of criterion-referenced decisions.

Table 1

Overview of the one facet pxi design with person (p) as the object of measurement and item (i) as the single facet.

Source	Number of Conditions Sampled	Associated Variance Component
person	n_p	σ_p^2
item	n_i	σ_i^2
error		σ_e^2

relative error = $\frac{\sigma_e^2}{n_i}$ G Coefficient = $\frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_e^2}{n_i}}$

absolute error = $\frac{\sigma_i^2}{n_i} + \frac{\sigma_e^2}{n_i}$ phi (ϕ) coefficient = $\frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_i^2}{n_i} + \frac{\sigma_e^2}{n_i}}$

GT provides two general purpose reliability coefficients as well as a coefficient or index for domain-referenced interpretations involving a fixed cut-off score (Brennan, 1983; Shavelson and Webb, 1981). We focus on the G coefficient for relative assessment, and the phi coefficient for absolute assessment. As illustrated in Table 1, the G coefficient for the single-facet situation involves the variance components for both the object of measurement and error. The phi coefficient, however, involves all three variance components. Item variability affects the absolute magnitude but not the relative placement of person's scores. Thus item variability contributes to error variability for absolute decisions but not for relative decisions.

Behavioral or knowledge measurements may be either norm- or criterion-referenced. However, this will seldom be the case with clinical

measurements. For example, to tell someone they are in the 90th percentile for blood pressure is meaningless!! This rank doesn't tell them what they need to know, they want to know the true value of their BP and are looking for it to be within the acceptable level for individuals in their age group.

The G coefficient and the phi coefficient have often been referred to as "reliability-like" coefficients, but O'Brien (1995) points out that the G coefficient is a reliability coefficient. That G coefficients are reliability coefficients is underscored by the fact that for the simple one-facet situation, there is an equivalence between reliability coefficients produced by CT and those produced by GT. In particular, the G coefficient of Table 1 is equivalent to Cronbach's alpha as well as Kuder-Richardson Formula 20 (Cronbach et al., 1972). Additionally, Cronbach et al. (1972) show the equivalence between results from GT and the Spearman-Brown adjustment for both the crossed design and the design having items nested within persons.

Multiple facets - the pxixo design.

One of GT's greatest and most touted strengths is that it can appropriately model measurement situations having multiple sources of error. For example, McGaghie et al. (1993) discuss the development of an instrument to measure attitudes toward pulmonary disease prevention. In reporting the reliability of the instrument, they report both test-retest and internal consistency reliabilities. This approach does not allow accurate assessment of the reliability of a measurement for persons (p) based on both multiple items (i) and occasions (o). The fully crossed design is denoted pxixo and has corresponding linear model

person-item-occasion score = grand mean + p main effect + i main effect + o main effect + pxi interaction effect + pxo interaction effect + ixo interaction effect + (error, pxixo interaction).

The fully crossed situation involves giving all the items to all the people on each occasion (Table 2). If different items were used on each occasion, then items would be nested within occasion and the design would be denoted px(i:o). Such a nested design can be appropriately modeled in GT but cannot provide separate estimates of the item main effect and ixo interaction variance components. Similarly, the pxi variance component will be confounded with the error and the three-way interaction.

Table 2

Overview of the two facet pxixo design with person (p) as the object of measurement and item (i) and occasion (o) as facets.

Source	Number of Conditions Sampled	Associated Variance Component
person	n_p	σ^2_p
item	n_i	σ^2_i
occasion	n_o	σ^2_o
person x item		σ^2_{pi}
person x occasion		σ^2_{po}
item x occasion		σ^2_{io}
error		σ^2_e

$$\text{relative error} = \frac{\sigma^2_{pi}}{n_i} + \frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_e}{n_i n_o} \qquad \text{G coefficient} = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{pi}}{n_i} + \frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_e}{n_i n_o}}$$

$$\text{absolute error} = \frac{\sigma^2_i}{n_i} + \frac{\sigma^2_o}{n_o} + \frac{\sigma^2_{pi}}{n_i} + \frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_{io}}{n_i n_o} + \frac{\sigma^2_e}{n_i n_o}$$

$$\text{phi } (\phi) \text{ coefficient} = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_i}{n_i} + \frac{\sigma^2_o}{n_o} + \frac{\sigma^2_{pi}}{n_i} + \frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_{io}}{n_i n_o} + \frac{\sigma^2_e}{n_i n_o}}$$

GT allows the researcher to examine the variance components associated with the object of measurement, facets, and all interactions. Each

of these variance components is then properly interpreted in the context of the measurement to be formed and the type of decision to be made. In this example, a person's score will be formed by averaging across all items and all occasions so that to estimate the score's error variance, each component is divided by the number of conditions of each facet that affects that component (Table 2). The variance of the relative error which is used in estimating the G coefficient includes the error variance component as well as all variance components associated with an interaction with persons. These are the only variance components affecting the relative placements of the total scores. The variance of the absolute error includes all facet main effect, all interaction, and the error variance components because all of these variance components contribute to fluctuations or variability in the absolute score. Because GT's reliability coefficients use either the variance of the relative error (G coefficient) or the variance of the absolute error (phi coefficient), these coefficients are correctly tailored to the form of the measurement, even if it is complex and involves summing over several different facets.

Organization as the object of measurement

An important strength of GT is the range of choices for an object of measurement (Brennan, 1983; Kane, 1993). This advantage may have particular relevance to health educators.

Golaszewski et al. (1990), note that practitioners and evaluators of health promotion and health education face the challenge of providing evidence of accountability and program efficacy by administering reliable measures to assess the impact of their initiatives, methods, and interventions. They apply the tools of CT to the assessment of the reliability of a worksite health promotion program. Their analysis, however, treats individuals as objects of measurement but for these programs, program customers were organizations comprised of many individuals. When computing an organization or aggregate-level score using the response of many individuals, individuals become a source of error much like items are when a score is obtained for an individual by summing responses to several items. Thus the reliability of individual scores may have little relationship to the reliability of the score for the organization.

O'Brien (1990) notes that conditions that result in high reliability for aggregate-level variables (such as an overall organization score) are not the same as the conditions that lead to high reliability for individual-

level scores. In fact, it is possible to have very high reliability for the individual-level variable but very low reliability when these individual-level variables are summed or averaged to obtain an aggregate-level variable. O'Brien (1990) considers the application of GT methodology to assessing the reliability of such aggregate-level variables in detail. The crux of his approach involves utilizing the power and flexibility of the linear models methodology that underlies GT to appropriately model a measurement situation having organization as the object of measurement and person as a facet.

A simple example involving an aggregate-level variable might include organizations (o), items (i), and persons (p) within organizations. Such a design might be denoted by (p:o)xi. That is all items are administered to all persons but persons are nested within organizations. This scenario makes sense if organization membership is mutually exclusive and the same set of items is administered to all of the sampled persons in each of the sampled organizations. Table 3 gives a brief breakdown of the estimable variance components and summarizes estimation of the G and phi coefficients when organization is the object of measurement. Due to the structure of the universes of admissible observations and of generalization, a randomly chosen person belongs to a single randomly chosen organization and no separate estimate of the variance component for person exists. Brennan (1975) uses GT methods to derive estimates of reliability coefficients both for an aggregate-level measurement (i.e., schools) and for persons nested within the aggregate. In measurement situations where aggregate-level measurements are of interest, it is possible for individual-level scores to have high reliability even if organization-level scores have low reliability. This might occur if, within each organization, person-to-person variability were large while overall differences between organizations were small. In this case, if overall differences among organizations were negligible it might not be possible to form meaningful organization scores for the purpose of ranking the organizations.

A fundamental flaw with CT coefficients is that, for many situations, the model on which they are based simply is not correct - items may vary in difficulty, multiple sources of error may be present, or something other than persons may be the object of measurement. A model must be appropriate or nearly correct to provide reasonably accurate insight into the modeled situation. As Kane (1993) notes, GT emphasizes designing data analysis to fit the data but CT applies a narrowly defined prespecified

Table 3

Overview of (p:o)xi design with organization (o) as the object of measurement and persons (p) and items (i) as facets. Persons are nested within organization with the same number n_p of person sampled from each organization for a total of $n_p x n_o$ persons.

Source	Number of Conditions Sampled	Associated Variance Component
organization	n_o	σ^2_o
item	n_i	σ^2_i
person: organization	$n_p(n_o)$	$\sigma^2_{p,op}$
organization x item		σ^2_{oi}
error		σ^2_e

relative error = $\frac{\sigma^2_{p,op}}{n_p} + \frac{\sigma^2_{oi}}{n_i} + \frac{\sigma^2_e}{n_p n_i}$

G coefficient = $\frac{\sigma^2_o}{\sigma^2_o + \frac{\sigma^2_{p,op}}{n_p} + \frac{\sigma^2_{oi}}{n_i} + \frac{\sigma^2_e}{n_p n_i}}$

absolute error = $\frac{\sigma^2_i}{n_i} + \frac{\sigma^2_{p,op}}{n_p} + \frac{\sigma^2_{oi}}{n_i} + \frac{\sigma^2_e}{n_p n_i}$

phi (ϕ) coefficient = $\frac{\sigma^2_o}{\sigma^2_o + \frac{\sigma^2_i}{n_i} + \frac{\sigma^2_{p,op}}{n_p} + \frac{\sigma^2_{oi}}{n_i} + \frac{\sigma^2_e}{n_p n_i}}$

model to the data whether that model is appropriate or not. In cases where both multiple items and multiple occasions are used in obtaining a score, this means that CT runs the data through several analyses and comes up with separate "reliability coefficients" which leave one wondering what the actual reliability of the measurement might be. Attempts to combine information from these analyses are likely to be misleading since it is not possible from these separate analyses to determine whether or how much of the variability due to different facets overlaps and how much addi-

tional error is actually due to interaction among the sources of error (Eason, 1989). GT, however, allows simultaneous consideration and modeling of all sampled sources of error and, from the model, estimates variance components. Estimated variance components can then be used to estimate the reliability of even a complex measurement involving summing over several facets.

GT: A Powerful Tool for Examining the Dependability of Measurements

GT provides a conceptual framework and the technical apparatus for examining the dependability of measurements. Within this conceptual framework, the distinct notions of a G study and a D study arise. G studies are associated with the development of a measurement procedure while D studies apply the procedure in practical terms. However, G studies vary in scope and may provide information that is the basis for a single D study or for several D studies. If the results of a G study show some facets to contribute little to the variability in a measurement, one may ignore or reduce the number of levels of that fact used in the D study. However, if a facet contributes a great deal of variability to the measurement, generalizability of D study measurements may be increased by increasing the number of levels of that facet used in the D study.

G studies may be designed for the purpose of gaining broad insight into a measurement situation. Such a G study may provide information for planning a number of different D studies. One of the great contributions of GT is that it places responsibility on the researcher to consider carefully the measurement process and to attempt to determine all facets that exist in both the universe of generalization and in the universe of admissible observations. A carefully considered G study no longer see measurements in the simplistic light of CT. For example, consider designing a G study to examine the reliability of blood pressure measurements. The researcher should consider all potential sources of error, including days, and time of day, as well as the environment in which the measurement is taken, measuring device, and measuring device operator. In a G study, analysis centers on those facets that are systematically represented by sampling multiple conditions of the facet. Facets not systematically represented may affect the measurement process but such facets cannot be explicitly considered in the analysis. Cronbach et al. (1972) emphasize the importance of implicit facets and their impact on the inter-

pretation of estimated coefficients.

Implicit facets appear in one of three ways. Implicit facets may be held constant throughout the G study, the condition of the implicit facet may vary throughout the G study without direct experimenter control, or conditions of an implicit facet may be confounded either with the object of measurement or with some other facet. While G coefficients are generic reliability coefficients (O'Brien, 1995), they are also specific with regard to facets that are held constant at the same condition throughout the study. For example, many studies are thought to include the single facet item but, in fact, occasion may be a specific implicit facet that can not be accounted for in the analysis. In this case, reliability coefficients computed for data gathered on a single occasion are specific to that occasion but generic with respect to items on that specific occasion.

Clear description of a G study allows those attempting to use G study information in planning a D study to know how G study information relates to their study. While implicit facets are not analyzed directly, their role in a G study should be reported because implicit facets have implications for the interpretation of study results and estimated variance components. G study reports should include results such as the numbers of conditions of each facet and estimates of variance components. Additionally, G study reports should include descriptions of: subjects' characteristics; the data collection; the nature of the conditions sampled; any conditions held constant; any conditions confounded with a facet; and any conditions intentionally allowed to vary at random without experimental control (Cronbach et al., 1972).

If the aim of a G study is to obtain broad insight into a measurement process, then a design that is as fully crossed as possible should be used. Crossed designs allow separate estimation of both main effect and interaction variance components. Information from a crossed design can be used in designing a nested D study. Nested designs however, do not allow separate estimation of some variance components and so may yield somewhat less information about the measurement process. The use of crossed designs for G studies has been widely advocated in the GT literature since these designs can be used as the basis for planning a variety of D studies. However, Cronbach et al. (1972) note that when components are to be confounded in the D study design then, a G study with similar confounding will yield more precise estimates for the planning of that particular D study.

Modeling BP

Physiological and biochemical functions such as BP tend to fluctuate around some central value. Ideally, a measurement process will give a reliable estimate of the individual's true central value. Obtaining a reliable estimate requires sampling frequently enough and including all major sources of variability in the sampling. Szalai et al. (1993) used GT to arrive at tables of measurement schemes that would yield "reliable" estimates of an individual's central characteristic BP. Their study included days (d), hours within days (h:d), and replications within hours within days (r:h:d). One implicit facet in their study was the particular device or instrumentation used to measure BP. Thus coefficients estimated using their data are specific to that instrumentation but generic with regard to days, hours, and replications.

A simple scenario might be designed to consider day to day variability in BP readings and to determine the reliability of a measurement obtained by averaging measurements taken on different days. Suppose that the G study information is obtained on the same days for all subjects (s), then the analysis of this situation looks very similar to that indicated in Table 1 but with the single facet days replacing the single facet items. As in the Szalai et al. (1993) study, instrumentation is a constant implicit facet. However, time of day also represents an implicit facet that may have an impact on the interpretation of the estimated variance components and therefore also the estimated reliabilities. If all readings in the study are taken at a particular time of day, then the facet time of day is a fixed implicit facet in the study. If, instead, the time of readings is allowed to vary without direct experimenter control, then time of day is still an implicit facet but it takes on a very different flavor. Which G study design is better depends on the ultimate D study situation. Suppose future D study readings are to allow time of day to vary randomly. If time of day has any impact on BP, then variance components estimated using data gathered at a fixed time may not be accurate estimates of the variance components for the D study design. It is likely that a G study with time of day fixed will underestimate the error variance component which will result in an optimistic estimate of the D study reliability. Sampling design and implicit facets have an impact on the interpretation and predictive utility of G study estimates.

Presumably, the design in Szalai et al. (1993) was nested because within each day included in the study, a different set of hours were ran-

domly chosen to be included in the study. Additionally, they had replications nested within hours within days so that their design was denoted by sx(r:h:d). We will consider and briefly compare two simpler designs, one crossed and one nested. The first is the sx(t:d) where t represents time (rather than hour) and the second is the sxdxt design. Note that in the nested design, because subject is crossed with t:d, each subject is sampled according to the same scheme. That is, on the same days and at the same time within each day. However, different times have been chosen within each day so that the sampling scheme varies from day to day. In the crossed design, a single set of random times has been selected and the sampling scheme stays the same from day to day. The crossed design may yield more information about the measurement process if there are daily cycles (i.e., time main effects) in BP. Table 4 summarizes key points of the sx(t:d) analysis while key points of the analysis for the sxtxd design are the same as for the pxixo design of Table 2. The crossed design allows separate estimation of seven different variance components while the nested design allows estimation of only five. In particular, the time main effect and time by day interaction from the crossed design are reflected in the single time within day component for the nested design. Additionally, the error for the nested design lumps the information for three components into one component. Information from a nested G study design does not provide the information needed to accurately plan a crossed D study because separate estimates of several components do not exist. Table 5 shows how to use estimates from the crossed design in planning a D study having the nested design. The crossed design provides information for a much broader array of possible D studies. However, estimates from the nested G study design will be more accurate for D studies having similar nesting of time within day.

Reliability as an attribute of the data

One implication of GT is that reliability is an attribute of the data, not of a given instrument or test (Thompson 1991, 1992). Both G studies and D studies involve sampling to obtain observations which are used to estimate variance components. Estimates of variance components are known to be rather poor when relatively few conditions are sampled. Because of this sampling error, estimates, even of the same underlying reliability, will vary from one application of a design to another. Furthermore, it is seldom the case that two data sets estimate exactly the same reliability. Universes of generalization differ from one study (especially

Table 4

Overview of the two facet sx(t:d) design with subject (s) as the object of measurement and day (d) and time (t) nested within day as facets. Each day, n_t times are randomly sampled to give a total of $n_t n_d$ times.

Source	Number of Conditions Sampled	Associated Variance Component
subject	n_s	σ^2_s
day	n_d	σ^2_d
time within day	$n_t(n_d)$	$\sigma^2_{t,dt}$
person x day		σ^2_{sd}
error		σ^2_e

$$\text{relative error} = \frac{\sigma^2_{sd}}{n_d} + \frac{\sigma^2_e}{n_d n_t} \qquad \text{G coefficient} = \frac{\sigma^2_s}{\sigma^2_s + \frac{\sigma^2_{sd}}{n_d} + \frac{\sigma^2_e}{n_d n_t}}$$

$$\text{absolute error} = \frac{\sigma^2_d}{n_d} + \frac{\sigma^2_{t,dt}}{n_t n_d} + \frac{\sigma^2_{sd}}{n_d} + \frac{\sigma^2_e}{n_d n_t}$$

$$\text{phi } (\phi) \text{ coefficient} = \frac{\sigma^2_s}{\sigma^2_s + \frac{\sigma^2_d}{n_d} + \frac{\sigma^2_{t,dt}}{n_t n_d} + \frac{\sigma^2_{sd}}{n_d} + \frac{\sigma^2_e}{n_d n_t}}$$

for studies by different investigators in different regions) to another - either the population of interest changes or the range of viable facet conditions changes so that the reliability calculated from one application of an instrument may not be an accurate estimate of the reliability for the next application of that instrument. Implicit facets change from one study to another. Something as simple as a change in an assistant administering a test or a change in the location at which a test is administered may cause a slight change in the reliability for the data being collected. Even the

Table 5

Use of estimates from sxtxd G study $(\hat{\sigma}^2_s, \hat{\sigma}^2_t, \hat{\sigma}^2_d, \hat{\sigma}^2_{st}, \hat{\sigma}^2_{sd}, \hat{\sigma}^2_{td}, \hat{\sigma}^2_e)$ *to plan an sx(t:d) D study. D study sample sizes may differ from original G study sample sizes. Denote D study sample size by* n'_d *and* n'_t.

D - Study sample sizes	n'_d , n'_t
Estimate of $\sigma^2_{t,dt}$	$\hat{\sigma}^2_t + \hat{\sigma}^2_{td}$
Estimate of σ^2_e (nested)	$\hat{\sigma}^2_{st} + \hat{\sigma}^2_e$
Estimate of relative error =	$\dfrac{\hat{\sigma}^2_{sd}}{n'_d} + \dfrac{\hat{\sigma}^2_{st} + \hat{\sigma}^2_e}{n'_d n'_t}$
G coefficient =	$\dfrac{\hat{\sigma}^2_s}{\hat{\sigma}^2_s + \dfrac{\hat{\sigma}^2_{sd}}{n'_d} + \dfrac{\hat{\sigma}^2_{st} + \hat{\sigma}^2_e}{n'_d n'_t}}$
Estimate of absolute error =	$\dfrac{\hat{\sigma}^2_d}{n'_d} + \dfrac{\hat{\sigma}^2_t + \hat{\sigma}^2_{td}}{n'_t n'_d} + \dfrac{\hat{\sigma}^2_{sd}}{n'_d} + \dfrac{\hat{\sigma}^2_{st} + \hat{\sigma}^2_e}{n'_t n'_d}$
Φ (phi) coefficient =	$\dfrac{\hat{\sigma}^2_s}{\hat{\sigma}^2_s + \dfrac{\hat{\sigma}^2_d}{n'_d} + \dfrac{\hat{\sigma}^2_t + \hat{\sigma}^2_{td}}{n'_t n'_d} + \dfrac{\hat{\sigma}^2_{sd}}{n'_d} + \dfrac{\hat{\sigma}^2_{st} + \hat{\sigma}^2_e}{n'_t n'_d}}$

occasion or the time period during which a study is conducted is specific to that data set and in that sense, the reliability coefficient for that data is specific for that time period.

D studies may generate large volumes of data at once or may generate information in a one-at-a-time fashion. For example, evaluation of health programs may provide an instance where many organizations are evaluated at once. In this case, D study data should be used to estimate the reliability of the D study measurements. For data generated in a one-at-a-time fashion, this may be difficult. For example, for the individual monitoring their own BP, there may be advantages to their understanding

what G studies have to say about the dependability of BP measurements. Not only can they use G study information to determine how to obtain a reliable assessment of their BP, but, an understanding of what GT has to say about the variability inherent in the BP measurement process may give them the information they need to put a single high or low measurement in perspective.

Concluding Remarks

The linear models machinery underlying GT provides a powerful and flexible framework for modeling a measurement situation. This framework is much richer than the limited framework of CT and allows GT to consider dependability issues for both norm- and criterion-referenced measures, for measures involving multiple sources of error, and for aggregate-level as well as individual-level measures. In addition, this technical machinery emphasizes the value of interpreting variance components to understand a measurement situation. We considered applying GT to some simple measurement situations. Reliabilities for measurements based on more complex measurement situations can be derived using linear models, variance components estimation, and appropriate interpretation of model components.

We have not considered the technical details of variance components estimation. While ANOVA estimators of variance components are frequently used, GT imposes no restriction on the mechanism used to estimate variance components and in some instances, other estimation methods may be preferable (Marcoulides, 1990; Shavelson and Webb, 1981). Khuri and Sahai (1985) provide a comprehensive review of variance components analysis including point estimation of variance components. Additionally, more advanced topics suggested by GT and its use of linear models include the use of regression estimates and confidence intervals in estimating true universe scores as alternatives to simply using observed scores (Cronbach et al., 1972), and multivariate generalizability allowing either the examination of the dependability of multiple scores simultaneously or the dependability of a composite score derived from multiple scores (Cronbach et al., 1972; Shavelson and Webb, 1981). G study variance component estimates may also be used to consider design optimization problems such as minimizing the number of observations per subject to achieve a specific generalizability coefficient (Sanders et al., 1989) or maximizing the coefficient of generalizability under resource

constraints (Marcoulides, 1997; Marcoulides and Goldstein, 1990; Sanders et al., 1991; Sanders, 1992). Other developments in GT appear in the literature including the possible link or integration of GT with validity theory (Kane, 1982; Shavelson and Webb, 1981).

In addition to the linear models technical machinery, GT provides a conceptual and contextual framework that places emphasis on asking the right questions. These questions arise naturally because it is the responsibility of the investigator to define the universes of admissibility and generalizability. Different universes of generalization have different universe-score variance, even when the procedure for obtaining a measurement appears, superficially, to be the same. Cronbach et al. (1972) note that error is often underestimated because investigators sampled from a universe narrower than that referred to in their theory. GT properly applied should require the investigator to at least question and think about the implicit aspects of his universe definition and how those aspects impact the interpretation and utility of variance component and reliability estimates.

By combining the linear models technical machinery with a strong conceptual framework GT provides a useful set of tools for examining the reliability of measurements. The health practitioner can use GT to assess the dependability of both norm- and criterion-referenced measurements used to evaluate health promotion programs and behavioral measurements. An understanding of GT also allows health educators to interpret information on the variability due to multiple sources of error from G studies of clinical measures.

Acknowledgements

The authors wish to thank Bernice L. Garrett for her assistance with typing and manuscript preparation and the College of Health and Social Services at New Mexico State University for financial assistance.

References

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice. 16*, 14-20.

Brennan, R. L. (1983). *Elements of Generalizability Theory.* IA: ACT Publications.

Brennan, R. L. (1975). The calculation of reliability from a split-plot factorial design. *Educational and Psycological Measurement, 35*, 779-788.

Brennan, R. L., and Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice. 14*, 9-12, 27.

Brown, M. T. (1992). Promoting use of generalizability theory: an impossible task? (A review of the book Generalizability theory: A primer). *Contemporary Psychology, 37*, 1327-1328.

Carswell, A., Dulberg, C., Carson, L., and Zgola, J. (1995). The functional performance measure for persons with Alzheimer disease: Reliability and validity. *Canadian Journal of Occupational Therapy, 62*, 62-69.

Cox, D. R. (1958). *Planning of Experiments*. New York: Wiley.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Croxton, F. E., and Cowden, D. J. (1955). *Applied general statistics*, (2nd ed.) Englewood Cliffs, N.J.: Prentice-Hall.

Eason, S. (1989). *Why generalizability theory yields better results than classical test theory*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Little Rock, AR, November 8-10, 1989) (ERIC Document Reproduction Service No. ED 314 434).

Golaszewski, T., Wassel, M. L., Yen, L., Lynch, W., and Vickery, D. M. (1990). The refinement of a worksite health promotion post-program evaluation instrument. *Health Education, 21*, 45-49.

Hatch, J. P., Prihoda, T. J., and Moore, P. J. (1992). The application of generalizability theory to surface electromyographic measurements during psycholophysiological stress testing: how many measurements are needed? *Biofeedback and Self-Regulation, 17*, 17-39.

Issac, S., and Michael, W. B. (1990). *Handbook in research and evaluation*. (2nd ed.) San Diego, CA.: Edits Publishers.

Kane, M. (1993). Review of the book Generalizability theory: A primer. *Journal of Educational Measurement, 30*, 269-272.

Kane, M. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125-160.

Kempthorne, O. (1957). *An introduction to genetic statistics*. New York: Wiley.

Khuri, A. I., and Sahai, H. (1985). Variance components analysis: A selective literature survey. *International Statistical Review, 53*, 279-300.

Kusumoto, F., Venet, T., Schiller, N. B., Sebastian, A., and Foster, E. (1995). Measurement of aortic blood flow by Doppler echocardiography: temporal, technician, and reader variability in normal subjects and the application of generalizability theory in clinical research. *Journal of the American Society of Echocardiography, 8*, 647-653.

Llabre, M. M., Ironson, G. H., Spitzer, S. B., Gellman, M. D., Weidler, D. J., and Schneiderman, N. (1988). How many blood pressure measurements are enough?: An application of generalizability theory to the study of blood pressure reliability. *Psychophysiology, 25*, 97-106.

Marcoulides, G. A. (1997). Optimizing measurement designs with budget constraints: The variable cost case. *Educational and Psychological Measurement, 57*, 808-812.

Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports, 66*, 379-386.

Marcoulides, G. A., and Goldstein, Z. (1990). The optimization of generalizability studies with resource constraints. *Educational and Psychological Measurement, 50*, 761-768.

McDaniel, A. M. (1994). Using generalizability theory for the estimation of reliability of a patient classification system. *Journal of Nursing Measurement, 2*, 49-62.

McGaghie, W. C., Boehlecke, B., DeVellis, B. M., Contreras, A., and Becker, M. (1993). Development of a measure of attitude toward pulmonary disease prevention. *Evaluation and the Health Professions, 16*, 106-118.

Norcini, J. J., and Day, S. C. (1995). Guidelines for interpretation of some common indicators of residency program performance. *The American Journal of Medicine, 98*, 285-290.

O'Brien, R. M. (1995). Generalizability coefficients are reliability coefficients. *Quality and Quantity 29*, 421-428.

O'Brien, R. M. (1990). Estimating the reliability of aggregate-level variables based on individual-level characteristics. *Sociological Methods and Research, 18*, 473-504.

Panse, V. G., and Sukhatme, P. V. (1954). *Statistical methods for agricultural workers*. New Delhi: Indian Council of Agricultural Research.

Roebroeck, M. E., Hariaar, J., and Lankhorst, G. J., (1993). The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. *Physical Therapy, 73*, 386-395.

Sanders, P. F. (1992). Alternative solutions for optimization problems in generalizability theory. *Psychometrika, 57*, 351-356.

Sanders, P. F., Theunissen, T. J. J. M., and Baas, S. M. (1991). Maximizing the coefficient of generalizability under the constraint of limited resources. *Psychometrika, 56*, 87-96.

Sanders, P. F., Theunissen, T. J. J. M., and Baas, S. M. (1989). Minimizing the number of observations: A generalization of the Spearman-Brown formula. *Psychometrika, 54*, 587-598.

Schaeffer, G. A., Carlson, R. E., and Matas, R. L. (1986). Assessing the reliability of criterion-referenced measures used to evaluate health-education programs. *Evaluation Review, 10*, 115-125.

Scheffè, H. (1959). *The analysis of variance*. New York: Wiley.

Searle, S. R. (1971). *Linear models*. New York: Wiley.

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

Shavelson, R. J., Webb, N. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133-166.

Shavelson, R. J., Webb, N. M., and Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*, 922-932.

Szalai, J. P., Reeves, R. A., and Katic, M. (1993). Reliability of blood pressure estimated through generalizability theory. *Pharmaceutical Medicine, 7*, 37-45.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70*, 434-438.

Thompson, B. (1991). Review of the book Generalizability theory: A primer. *Educational and Psychological Measurement, 51*, 1069-1075.

VanLeeuwen, D. M. (1997). Assessing reliability of measurements with generalizability theory: an application to inter-rater reliability. *Journal of Agricultural Education. 38*, 36-42.

Watkins, B., Darrah, J., and Pain, K. (1995). Reliability of passive ankle dorsiflexion measurements in children: comparison of universal and biplane goniometers. *Pediatric Physical Therapy, 7*, 3-8.

Webb, N. M., Rowley, G. L., and Shavelson, R. J. (1988). Methods, plainly speaking: using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development. 21*, 81-90.

JOURNAL OF OUTCOME MEASUREMENT®, 2(4), 326 - 337

The Job Responsibilities Scale: Invariance in a Longitudinal Prospective Study

Larry H. Ludlow
Boston College

Mary E. Lunz
American Society of Clinical Pathologists

The purpose of the present analysis was to determine the degree of invariance of the Job Responsibilities Scale from 1993 to 1995. Factor analyses were conducted on both year's data and nearly identical factor patterns were found. Rasch rating scale analyses were conducted and nearly identical pairs of item estimates were found. These results suggest that even though the overall frequency of performance on some medical technology laboratory tasks increased from 1993 to 1995, the relationships among the tasks themselves remained the same (invariant). This conclusion allows for a description of what it means to increase in level of personal job responsibility from year-to-year. In addition, these results suggest that at the conclusion of this prospective study it may be possible to objectively define the typical career mobility pattern of entry level medical technologists.

Correspondence concerning this article should be addressed to Larry H. Ludlow, Boston College, School of Education, Chestnut Hill, MA, 02467; e-mail: ludlow@bc.edu.

In 1993 the Research and Development committee of the Board of Registry of the American Society of Clinical Pathologists (ASCP) began a ten year longitudinal study of medical technologist jobs and related issues. The overall purpose was to follow a cohort of technologists from the year of certification across a ten year period, in order to better understand their job responsibilities and the social issues surrounding those responsibilities.

The design strategy was a series of surveys, administered according to a prescribed plan during the ten year period of the study. The survey for the first year of data collection included a Job Responsibilities Scale (JRS) which was designed and field tested by the members of the Board of Registry Research and Development committee, a group that included both laboratory science and statistics professionals. The plan was to administer the JRS in 1993 and again in 1995.

It was anticipated that the more complex job responsibilities would be performed more frequently by the participants as they acquired more experience as laboratory professionals. That is, an individual's personal level of job responsibility was expected to change over time. In order to measure this change, however, it was expected that the relationships among the tasks themselves would remain the same. Some tasks would be performed more or less frequently over time but the relative level of responsibility demanded on those tasks was not expected to change. These expectations are addressed in this paper through analyses conducted to determine the degree of item invariance on the JRS from 1993 to 1995.

Method

Sample

A cohort of 2000 individuals who applied to take the 1993 certification examination was selected to represent all routes of entry into the certification examination and, subsequently, the field of laboratory medicine. All of the respondents had completed a bachelors degree and were qualified for the examination under an education or experience route. Most of the respondents had recently completed their degrees, and 1993 was their first year as a medical technologist. Of the original cohort, 1063 complete surveys were available for the first year (1993). The same JRS was sent in 1995 but only 665 complete surveys were available for analysis. Although the attrition was large, the samples are comparable from a demographic and geographic perspective.

Instrument

The 30-item Job Responsibilities Scale was developed by the Board of Registry Research and Development committee which included laboratory science and research methodology experts. Respondents were asked to rate each item based on the frequency of performance during the calendar year 1993 and 1995 respectively. The items on the JRS covered two major types of responsibilities including: 1) core job responsibilities which are tasks such as "perform routine and specialized lab tests," "correlate abnormal values with disease states," and "maintain confidentiality of results;" and 2) more advanced technical and management responsibilities including items such as "evaluate instruments for use in the lab," "train lab personnel," "establish technical procedures," "supervise personnel" and "supervise projects." In order to complete these responsibilities in the laboratory, the technologist must have the requisite technical skills; a broad knowledge base; judgment, analytical decision making, and management skills; communication training; and professional experience.

In 1993, the first administration, respondents were asked to rate each task based on the frequency it was performed on a 4 point rating scale: frequently = 4 points, sometimes = 3 points, rarely = 2 points, and never = 1 point. Frequency of task performance scores were the sum of the points from the rating scale for each of the job responsibilities. Exactly the same scale and directions were used at the second administration in 1995. Further detail on the sample, instrument, and methodology may be found in Ludlow (in press).

Results

Comparison of 1993 and 1995 factor analysis solutions

The same statistical procedures were carried out on both data sets. The Cronbach alpha was identical (α =.88). Both determinants were non-zero. Both Bartlett's test of sphericity were statistically significant. Both Kaiser-Meyer-Olkin measures of sampling adequacy were "marvelous" (>.90). The same common factor analyses were performed with varimax rotations, although oblique rotations did reveal correlations between the final factors. Finally, a two factor solution was accepted for each data set. The eigenvalues and percent of variance accounted for were practically identical (see Table 1), and the interpretation of each solution was the same.

Figure 1 presents the two factor varimax plot for 1993. Factor I contains tasks of the nature: evaluate new instruments, purchase reagents,

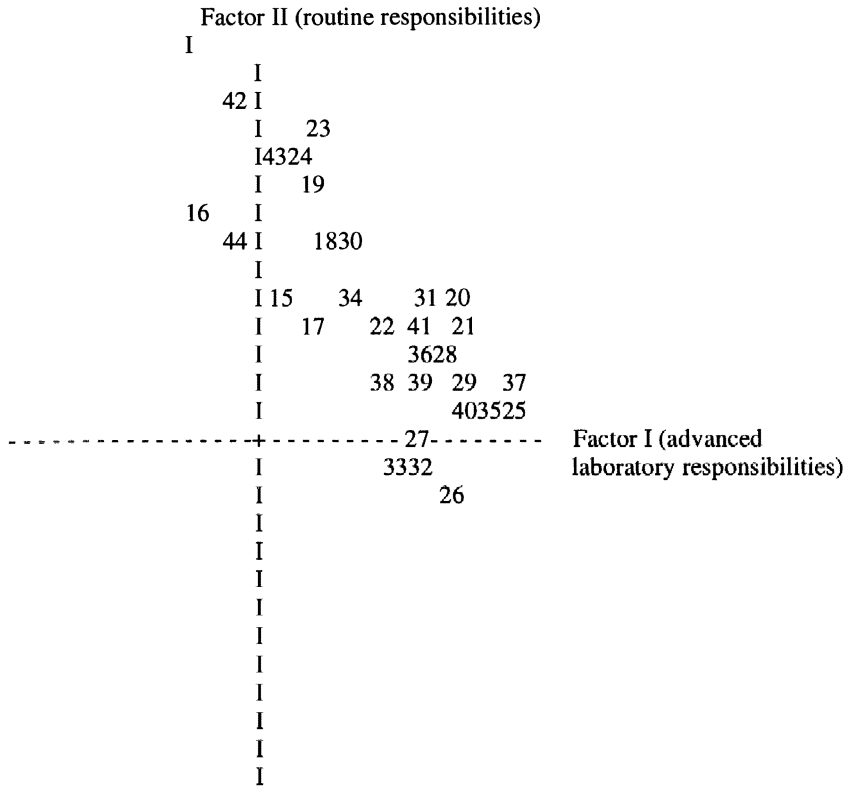


Figure 1. Varimax rotation for the two factor solution for 1993

participate in research activities, train laboratory personnel, present lectures, work on legislative activities, establish technical procedures, and supervise laboratory projects. This factor is labeled "advanced laboratory job responsibilities".

Factor II is defined by tasks of the nature: collect and prepare specimens, perform routine laboratory tasks, recognize a problem in quality control results, recognize normal and abnormal values, and maintain confidentiality of test results. This factor is labeled "routine responsibilities".

The two factor varimax plot for the 1995 data is presented in Figure 2. The patterns for 1993 and 1995 are strikingly similar. Essentially the same responsibilities load highest on Factor 1 and Factor 2.

The similarity between the two solutions is summarized in Table 1. The only tasks that shift their dominant factor loadings from 1993 to 1995 are "Q34", "Q30", and "Q18". Q34 asks how often respondents partici-

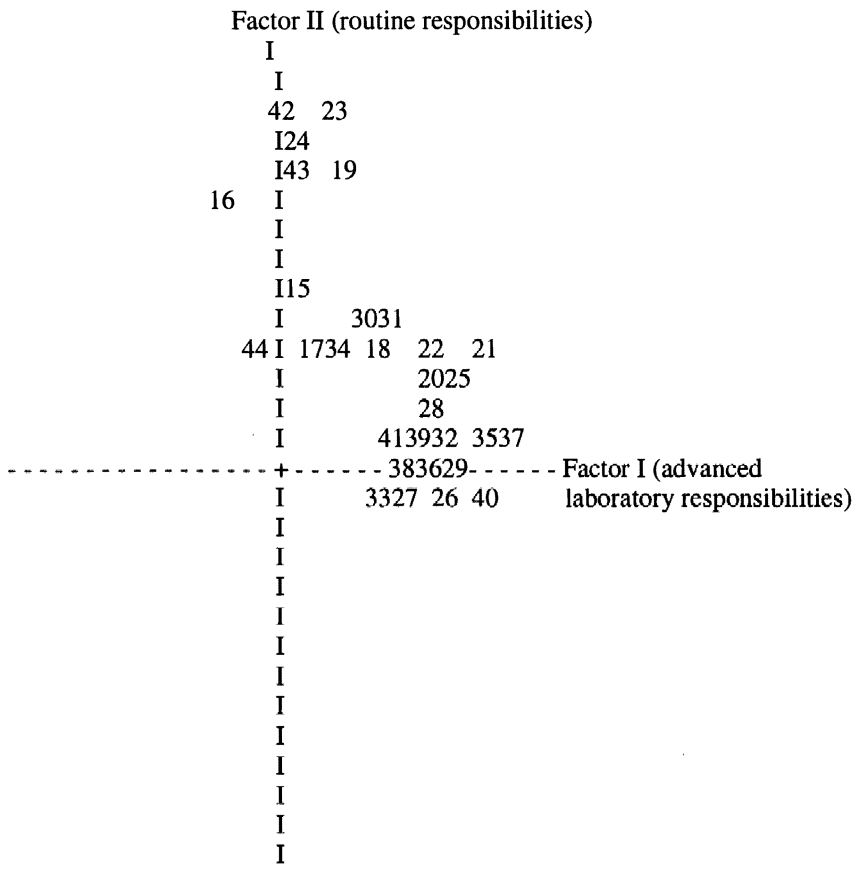


Figure 2. Varimax rotation for the two factor solution for 1995

pated in continuing education. This question was associated with advanced responsibility tasks in 1993 but became somewhat more routine by 1995. In contrast, Q30 (communicate technical information to medical and lay persons) and Q18 (perform quality assurance activities) became more closely associated with advanced responsibility tasks by 1995.

Cureton and D'Agostino (1983) suggest an index that is useful for comparing factors extracted from two data sets. Their coefficient of congruence (*CC*) between comparable factors is

$$CC = \frac{\sum l_1 l_2}{\sqrt{(\sum l_1^2)(\sum l_2^2)}},$$

Table 1

Varimax Solutions for the Job Responsibilities Scale: 1993 and 1995

		1993		1995	
ITEM NUM	ITEM NAME	FACTOR 1	FACTOR 2	FACTOR 1	FACTOR 2
23	Q37	.797	.070	.796	.049
11	Q25	.783	.061	.728	.130
21	Q35	.741	.064	.742	.045
26	Q40	.690	.065	.754	-.006
7	Q21	.660	.195	.657	.224
6	Q20	.650	.210	.668	.176
15	Q29	.642	.087	.609	.061
14	Q28	.637	.137	.689	.099
12	Q26	.605	-.069	.610	-.001
18	Q32	.566	-.010	.607	.011
25	Q39	.564	.114	.567	.026
22	Q36	.540	.171	.526	.034
27	Q41	.532	.195	.502	.060
17	Q31	.529	.217	.446	.276
13	Q27	.526	.008	.485	-.018
19	Q33	.498	-.022	.418	.008
24	Q38	.422	.100	.474	.046
8	Q22	.410	.223	.486	.192
20	**Q34**	**.311**	.289	.206	**.215**
28	Q42	-.052	.691	-.006	.694
9	Q23	.245	.642	.187	.660
10	Q24	.117	.588	.068	.648
29	Q43	.085	.560	.089	.574
5	Q19	.151	.557	.171	.591
2	Q16	-.161	.432	-.293	.519
30	Q44	-.067	.398	-.101	.238
4	**Q18**	.229	**.385**	**.309**	.228
16	**Q30**	.304	**.372**	**.391**	.275
1	Q15	.065	.292	.069	.337
3	Q17	.193	.210	.156	.229
Eigenvalue		7.67	2.45	7.50	2.58
% variance		25.60	8.20	25.00	8.60

Note: Tasks in bold-face changed their dominant factor loading positions.

where r_1 and r_2 are the corresponding rotated factor loadings from the two solutions. They suggest a guideline of CC >= .90. The CC for Factor 1 was .93 and the CC for Factor 2 was .85. Certainly Factor 1 meets the criteria while Factor 2 reflects the previously discussed slight shift in frequency of performance for a few tasks from 1993 to 1995. It seems reasonable to expect that as the 10 year study progresses, more of these shifts will occur.

Comparison of 1993 and 1995 Rasch IRT solutions

The Rasch rating scale model (Wright and Masters, 1982, p. 49) was applied to these data. This model specifies the probability of person n responding in category x to item i as

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} \left[\beta_n - \left(\delta_i + \tau_j\right)\right]}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} \left[\beta_n - \left(\delta_i + \tau_j\right)\right]},$$

where $x = 0,1, \ldots m$ rating scale categories.

With this model, a respondent's "level of responsibility attained" parameter (β_n), a set of scoring category "frequency performed" parameters (τ_j), and an item's "level of responsibility required" parameter (δ_i) are estimated. The 1993 solution is displayed as a "variable map" in Figure 3.

The X's to the left of the vertical line represent the individual medical technologist locations ($\hat{\beta}$) with regard to their overall level of responsibility on the JRS. Persons performing tasks requiring advanced levels of responsibility are located at the top of the map, persons with routine responsibilities are in the lower section. To the right of the vertical line are the locations of the job responsibility items ($\hat{\delta}_i$). The tasks requiring advanced levels of responsibility (Factor I) are in the upper region of the map while the routine tasks (Factor II) are in the lower region. The mean level of responsibility for the technologists was ($\hat{\beta}$ = -.24). This indicated that these laboratory technologists, overall, perceived that the task responsibilities they performed most frequently were mostly of a routine nature. All estimates are reported in the logit metric (see Ludlow and Haley, 1995).

The same rating scale analysis was performed upon the 1995 data. A statistical comparison of the two solutions reveals several similarities. The "person separation indices" were 2.7 (1993) and 2.8 (1995). The "item separation indices" were 28.9 (1993) and 21.7 (1995). The category threshold estimates were $\hat{\tau}_1 = -0.54, \hat{\tau}_2 = -0.22, \hat{\tau}_3 = 0.76$ (1993) and $\hat{\tau}_1 = -0.52, \hat{\tau}_2 = -0.15, \hat{\tau}_3 = 0.67$ (1995). Finally, a visual comparison of Figure 3 and Figure 4 reveals similar distributions of person responsibility estimates and a similar hierarchical ordering to the tasks.

The 1995 task estimates appear somewhat smoother in their distribution. That smoothness is attributed to the fact that there is now a greater

Logit position	Technologists (N=1063)	Tasks (L=30)	Task description
2.9			
2.8	X	(ADVANCED)	
2.7			
2.6			
2.5		Q33	legislation
2.4			
2.3			
2.2	X		
2.1			
2.0			
1.9	XX		
1.8	X	Q32	lectures
1.7	X		
1.6	XXX	Q26	purchasing
1.5	X	Q36, Q40	evaluation, super
1.4	X	Q25, Q37, Q39	eval, procedures,
1.3	XX	Q29	utilization
1.2	XX		
1.1	XXXXXXX		
1.0	XXXX		
0.9	XXXXXXXXXX	Q27, Q35	research, writing
0.8	XXXX	Q28, Q41	selection, commi
0.7	XXXXXXXXXXXX	Q21, Q38	evaluation, imple
0.6	XXXXXXXXXXX		
0.5	XXXXXXXXXXXXXXXXXXX		
0.4	XXXXXXXXXXXXX	Q20	implementation
0.3	XXXXXXXXXXXXXXXXXXXX		
0.2	XXXXXXXXXXXXXXXXXXXXX		
0.1	XXXXXXXXXXXXX	Q22, Q31	evaluation, traini
0.0	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX		
-0.1	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX		
-0.2	XXXXXXXXXXXXX		
-0.3	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX		
-0.4	XXXXXXXXXXXXXXXXX		
-0.5	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX		
-0.6	XXXXXXXXXXXXX	Q15, Q34	preparation, parti
-0.7	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX		
-0.8	XXXXXXXXXXXXXXXX		
-0.9	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	Q17, Q23, Q30	perform, trblshoo
-1.0	XXXXXXXXXXXXXXXX		
-1.1	XXXXXXXXX		
-1.2	XXXXXXXXXXX		
-1.3	XXXXXXXXXXXXXXXX	Q18	perform
-1.4	XXXXXXX	Q19, Q43	recognize, correl
-1.5	XXX	Q24	perform
-1.6	XXX		
-1.7	XX		
-1.8	XX		
-1.9			
-2.0	X		
-2.1	X		
-2.2			
-2.3			
-2.4			
-2.5			
-2.6			
-2.7			
-2.8			
-2.9			
-3.0		Q16	perform
-3.1		Q42	recognize
-3.2			
-3.3			
-3.4			
-3.5			
-3.6			
-3.7			
-3.8			
-3.9			
-4.0			
-4.1			
-4.2			
-4.3		Q44	confidentiality
		(ROUTINE)	

Figure 3. Variable map showing positions of people and tasks on the Job Responsibilities Scale for 1993.

Note: When the score group positions are closer than one tenth of a logit score groups are combined: X=2 persons in this map.

Logit position	Technologists (N=664)	Tasks (L=30)	Task description
4.8	X		
	this gap was removed to enable printing on one page-no persons or items were within this section		
3.1			
3.0		(ADVANCED)	
2.9			
2.8			
2.7		Q33	legislation
2.6			
2.5			
2.4			
2.3	X		
2.2	X		
2.1	X		
2.0	XX		
1.9	X		
1.8	XXX		
1.7			
1.6	XXXX	Q36	evaluation
1.5	XXXXX	Q32	lectures
1.4	XXXXXXX		
1.3	XXXXXXXXXX	Q29	utilization
1.2	XXXXXXX	Q26	purchasing
1.1	XXXXXXXXXXX	Q25 Q37	eval, procedures
1.0	XXXXXXXXXXX	Q27 Q40	
0.9	XXXXXXXXXXXXXXXXXX	Q38	implementation
0.8	XXXXXXXXXXXXXXX	Q28 Q39	selection, supaervisio
0.7	XXXXXXXXXXX	Q41	committees
0.6	XXXXXXXXXXXXXXXXXXXXX	Q35	writing
0.5	XXXXXXXXXXXXXX	Q21	evaluation
0.4	XXXXXXXXXXXXXXXXXXXXXXX		
0.3	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	Q20	implementation
0.2	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX	Q22	evsluation
0.1	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX		
0.0	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX		
-0.1	XXXXXXXXXXXX		
-0.2	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	Q31	training
-0.3	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	Q15	preparation
-0.4	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX		
-0.5	XXXXXXXXXXXXXXXXXXXXXXXXX	Q34	participation
-0.6	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	Q30	communicate
-0.7	XXXXXXXXXXXXXXXX		
-0.8	XXXXXXXXXXXXXXXXXXXXXX	Q17	perform
-0.9	XXXXXXXXXXXXXXXX	Q23	troubleshoot
-1.0	XXXXXX	Q18 Q43	perform, correlate
-1.1	XXXXXXXX		
-1.2	XXXXXXXXXXX	Q19	recognize
-1.3	XXXX	Q24	perform
-1.4			
-1.5	X		
-1.6	X		
-1.7	X		
-1.8	X		
-1.9			
-2.0	X		
-2.1			
-2.2	X		
-2.3			
-2.4			
-2.5			
-2.6			
-2.7		Q16 Q42	perform, recognize
-2.8			
-2.9			
-3.0			
-3.1			
-3.2			
-3.3	X		
-3.4			
-3.5			
-3.6			
-3.7			
-3.8		Q44	confidentiality
-3.9			
-4.0			
-4.1	X	(ROUTINE)	

Figure 4. Variable map showing positions of people and tasks on the Job Responsibilities Scale for 1995

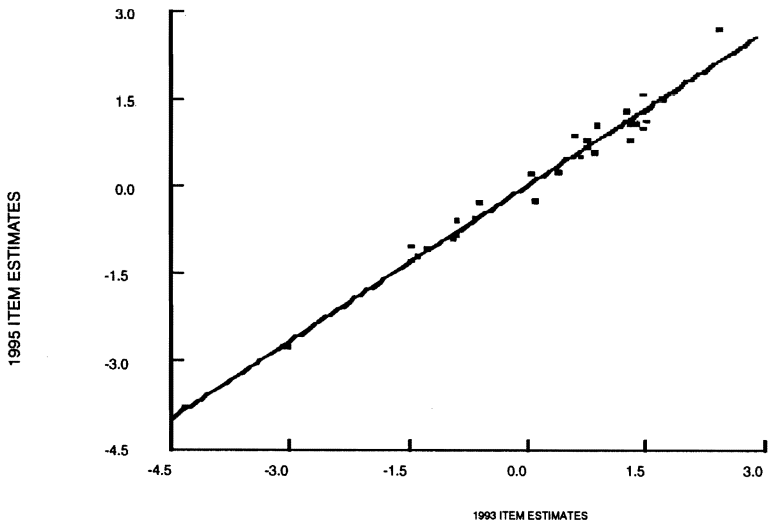
Note: When the score group positions are closer than one tenth of a logit score groups are combined: X=2 in this map.

range of tasks that the technologists are performing on a more regular basis[1]. This is particularly evident for the clusters of tasks located at about (1.5) and (-1.2) logits in 1993. By 1995 the tasks in those regions had become more clearly differentiated from one another because more technologists performed advanced tasks more frequently.

The mean person estimate in 1995 is ($\hat{\beta} = 0.01$). This represents an increase in personal job responsibility that roughly corresponds to literally moving up the JRS from "preparing specimens" and "participating in continuing education" (1993) to "implementing new test procedures" and "evaluating computer data and problems" (1995).

The worst fitting[2] task in 1993 was Q15: collect and prepare specimens. This task was again the worst fitting one in 1995. The interpretation for the misfit remains the same. There were numerous technologists with relatively high levels of responsibility who simply did not frequently perform this task. Unfortunately, from a Rasch perspective, these higher-level technologists were expected to perform the task.

Another way to investigate scale invariance is to use the 1993 task estimates as predictors of the 1995 estimates. A simple regression was performed and the results are presented in Figure 5. The regression is



($r = .99$, $b = .88$, $p<.0004$)

Figure 5. Regression of 1995 item estimates on 1993 item estimates.

a near identity and the standardized regression coefficient (beta=.99) is as high as one could reasonably seek.

The one obvious shift in the pairs of estimates lies in the upper right-hand region of the figure[3]. This task is Q33: work with legislative activities. Even the few technologists who were surprisingly performing this activity in 1993 were no longer performing it as frequently during 1995. This task, both conceptually and statistically, is the furthest removed from routine laboratory tasks.

Discussion

The purpose of these analyses was to examine the extent to which the tasks of the JRS remained invariant from 1993 to 1995. Regardless of the approach taken to investigate this problem the results were consistent. The factor analyses revealed that the tasks were comprised of two sets of related laboratory activities: relatively routine versus advanced responsibility tasks. The IRT analyses revealed a continuum progressing from the routine tasks upwards through advanced responsibility tasks as the technologists gained more experience.

The overall career mobility pattern was for technologists to increase the number of job responsibilities performed more frequently during the three year period. Although there are some specific exceptions, most technologists continued to perform the routine level responsibilities frequently but also added the advanced technical and management responsibilities to the list of more frequently performed tasks. Thus, the results indicate that individuals in this profession perform a core of responsibilities that remains fairly consistent over time even while they add more complex or advanced responsibilities as they gain experience. However, other issues such as job changes, promotion, transfer, geographic moves, and hospital organizational characteristics all interact with how frequently medical technologists perform specific job responsibilities.

Technologists have several alternative paths for advancement. They may continue to perform laboratory tests (e.g., blood glucose, PSA) for their entire careers, or they may choose to work toward a promotion which will yield more responsibility and the opportunity to assume more advanced responsibilities. The ramification of these choices is that some technologists are likely to continue to perform routine responsibilities during their entire careers. In fact, individuals who do achieve promotion do assume significantly more job responsibility for advanced manage-

ment and technical responsibilities (Lunz, Harmening, and Castleberry, submitted), as well as continuing to perform some of the same routine job responsibilities.

The results from this study are important from a practical measurement perspective because a shift in the definition of the JRS would have confounded the interpretation of what it means to move upward in responsibility over the first 10 years in the career of a laboratory technologist. Furthermore, these results are heartening because they provide preliminary evidence that at the conclusion of the current 10 year prospective study it should be possible to objectively define the first 10 years of the typical job mobility pattern to be expected of most entry-level medical technologists.

Footnotes

[1] If the reader were to superimpose the two maps such that corresponding logit estimates were aligned, then this spreading out of the 1993 clusters would be even more obvious.

[2] Fit was computed using the variance-weighted t statistic of Wright and Masters (1982, p. 100).

[3] Although, the shift in item estimates appears dramatic in this plot, the standardized difference statistic z was only (-1.87) (Wright and Masters, 1982, p. 115).

Acknowledgement

The graphics assistance of Dorothy Cochran is gratefully acknowledged.

References

Cureton, E.E. and D'Agostino, R.B. (1984). *Factor analysis: An applied approach.* LEA: Hillsdale, NJ.

Ludlow, L.H. (in press). Scale structure: An integrative data analysis approach.

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

Lunz, E., Harmening, D. and Castleberry, B.M. (submitted). Effects of reducing staff in the laboratory on task responsibilities, job satisfaction and wages.

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

JOURNAL OF OUTCOME MEASUREMENT®, 2(4), 338 - 350

Identifying Measurement Disturbance Effects Using Rasch Item Fit Statistics and the Logit Residual Index

Robert E. Mount
Dallas Public Schools

Randall E. Schumacker
University of North Texas
University of North Texas Health Science Center

A Monte Carlo study was conducted using simulated dichotomous data to determine the effects of guessing on Rasch item fit statistics (weighted total, unweighted total, and unweighted between fit statistics) and the Logit Residual Index (LRI). The data were simulated using 100 items, 100 persons, three levels of guessing (0%, 25%, and 50%), and two item difficulty distributions (normal and uniform). The results of the study indicated that no significant differences were found between the mean Rasch item fit statistics for each distribution type as the probability of guessing the correct answer increased. The mean item scores differed significantly with uniformly distributed item difficulties, but not normally distributed item difficulties. The LRI was more sensitive to large positive item misfit values associated with the unweighted total fit statistic than to similar values associated with the weighted total fit or unweighted between fit statistics. The greatest magnitude of change in LRI values (negative) was observed when the unweighted total fit statistic had large positive values greater than 2.4. The LRI statistic was most useful in identifying the linear trend in the residuals for each item, thereby indicating differences in ability groups, i.e. differential item functioning.

Requests for reprints should be sent to Randall E. Schumacker, College of Education, Technology and Cognition, Mathews Hall 304, University of North Texas, Denton, Texas 76203-1337, e-mail: rschumacker@unt.edu

After a unit or course of study, teachers usually assess the acquisition of knowledge based on the endorsement of items on a test. These items are often placed on a test without the knowledge of their true difficulty or function within different subgroups. It is also thought that the items are content specific and measure a single dimension or construct validly. We know that when a person responds to an item, that response is influenced by characteristics of the person that is independent of the items, and that the items possess certain qualities that are independent of the person. Given the diverse nature of individuals in a classroom, the accuracy of measurement is tantamount to understanding item function with the diverse nature of individual differences. This study investigated the effects of guessing on Rasch item fit statistics (weighted total, unweighted total, and unweighted between fit statistics) and the Logit Residual Index (LRI) as the probability of guessing the correct answer increased (0%, 25%, and 50%) and the usefulness of these statistics when applied to data from simulated teacher-made tests.

It has been shown that the Rasch model allows for the accurate measurements of individual differences on a true linear scale (Rasch, 1960/1980). No other mathematical model allows for the independent estimation of person ability measures and item difficulty calibrations (Anderson, 1973; Barndorff-Nielsen, 1978; Rasch, 1961; Wright and Stone, 1979). The logistic function in the Rasch model provides for both linearity of scale and generality of measure (Wright & Stone, 1979). Georg Rasch called this particular characteristic "specific objectivity." Therefore, accurate estimates of person ability and item difficulties are possible, yet measurement disturbances must be identified and taken into consideration.

Measurement Disturbances

Measurement disturbances are conditions that interfere with the measurement of some underlying psychological construct. Thorndike (1949) developed a list of possible disturbances to the measurement process. Smith (1985) later classified measurement disturbances into three general categories: (a) disturbances that are the results of characteristics of the person that are independent of the items, (b) disturbances that are the interaction between the characteristics of the person and the properties of the items, and (c) disturbances that are the results of the properties of the items that are independent of the characteristics of the person. The classification of measurement disturbances is important in that the source of the measurement disturbance dictates the techniques necessary to de-

tect its presence. Disturbances that are characteristics of the person and independent of the items include, but are not limited to (a) start-up, (b) plodding, (c) cheating, (d) illness, (e) boredom, and (f) fatigue. Measurement disturbances associated with the interaction of the person and the properties of the items are (a) guessing, (b) item content, (c) item type, and (d) item bias. With the Rasch model, only two conditions determine the outcome of the interaction between the person and any item on a test: (a) the amount of the trait possessed by the person, and (b) the amount of the trait necessary to provide a certain response to a given stimulus (Smith, 1991a). These conditions are commonly referred to as *person ability* and *item difficulty*. Any other conditions that influence outcomes are considered measurement disturbances. Glaser (1949, 1952) and Mosier (1941) felt that a person would exhibit consistently correct answers to relatively easy items, consistently incorrect responses to difficult items, and inconsistent responses to items centered on their ability level. Since inconsistent responses could be associated with measurement disturbances, Thurstone and Chave (1929) believed that some criterion should be established so that inconsistent responses could be eliminated.

Rasch Fit Statistics and the Logit Residual Index (LRI)

Rasch parameter estimates were found to be consistent, efficient, sufficient, and unbiased (Anderson, 1973; Andrich, 1988; Habermann, 1977; Wright, 1977; Wright & Stone, 1979). Consequently, analysis of fit for the entire response matrix does not require additional information beyond the item difficulties and person responses. This allows for the creation of subgroups (persons) that can be used to test the invariance of the item difficulty parameters and/or differences across subgroups. The interpretation of the outcomes becomes more useful when the observed fit statistics are based on some characteristics of the persons (age, gender, native language, or ethnic origin). Smith (1988b) performed several simulations to assess the distributional properties of the weighted and unweighted item between fit statistics. These simulations involved 10 replications of 1,000 persons taking a 20-item test, with the item difficulties uniformly distributed from -1 to +1 logits. The results showed that, as the number of ability groups increased, the mean and standard deviation of the transformed fit values approached the hypothesized values of 0,1. Additional simulations studied the effect of increasing the number of persons and number of items, varying the dispersion of item difficulties, and varying the offset between the mean of the item and person distribu-

tions. The results indicated that, within the ranges studied, varying these factors had little effect on the distribution of the transformed fit values. Thus, there appears to be no reason to develop correction factors such as those developed for the weighted and unweighted total fit statistic to correct for the influence of these factors on the distribution and Type I error rate of the item between fit statistics (Smith 1991b).

Smith (1988a, 1991b) and Smith and Hedges (1982) also studied the power of the total and between item fit statistics to detect two types of measurement disturbances, item bias and guessing. These studies found that the weighted total, unweighted total, and unweighted between fit statistics were capable of detecting different types of measurement disturbances. The between fit statistic was more efficient at detecting item bias than either the unweighted or weighted total fit statistic. The unweighted and weighted total fit statistics were more sensitive to disturbances such as guessing and start-up. The primary difference between the two statistics is that the unweighted version is based on the sum of the standardized residuals, whereas the weighted version is based on the sum of the standardized residuals that have been weighted by the information function.

IPARM (Item and Person Analysis with the Rasch Model), a computer software program introduced by Smith (1991b) performs Rasch item and person analyses from dichotomous and rating scale data. The major advantage of the program is that it constructs between fit statistics based on characteristics of the persons for item analysis and properties of the items for person analysis. It also provides between fit statistics (unweighted version) for biographical sub-populations. When demographic data are used in the analysis to create subgroups (sex, race, and age), the resulting statistics will give an indication of the presence of bias, or differential item familiarity in the response patterns for the items. The software first calculates the item mean squares associated with each Rasch fit statistic, then converts them to their unit normal fit statistic with the following cube root transformation;

$$t = (V^{1/3} - 1)\left(\frac{3}{S}\right) + \left(\frac{S}{3}\right),$$

where V, the mean square, and S, the standard deviation, are the values associated with the mean square under consideration (Smith, 1991b; Wright and Masters, 1982). The resulting Rasch item fit statistics have expected values of mean = 0 and standard deviation =1.

The weighted mean square item (WMS_i) is calculated as,

$$WMS_i = \frac{\sum_{n=1}^{N}(X_n - P_n)^2}{\sum_{n=1}^{N} W_n},$$

where P_n is the probability of a correct answer (predicted response) that can be calculated as

$$P_{ij} = \frac{\exp(b_j - d_i)}{1 + \exp(b_j - d_i)},$$

where b_j is the ability measure for persons in score group j, d_i is the item difficulty (Smith, 1991b, p. 153) and W, the weighting function, can be calculated as

$$W = [P(1-P)].$$

The standard deviation associated with the weighted total mean square can be calculated as

$$S[MS(WT)_i] = \frac{\left[\sum_{n=1}^{N} W_{ni} - 4\sum_{n=1}^{N} w_{ni}^2\right]^{1/2}}{\sum_{n=1}^{N} W_{ni}}.$$

The unweighted mean square item (UMS_i) is defined as

$$UMS_i = \frac{1}{N}\sum_{n=1}^{N}\left[\frac{(X_n - P_n)^2}{P_n(1-P_n)}\right] = \frac{1}{N}\sum_{n=1}^{N} Z_n^2,$$

where N is the number of persons, X_n is the observed response, and P_n is the response predicted from the logit difficulty of the item and the logit ability of the person. The standard deviation of the unweighted total mean square item can be found as

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

The unweighted between mean square item ($UBMS_i$) is defined as

$$\text{UBMS}_i = \frac{1}{(J-1)} \sum_{j=1}^{J} \frac{\sum_{n \in j}^{N_j} (X_n - P_n)^2}{\sum_{n \in j}^{N_j} P_n(1-P_n)},$$

where J is the number of score groups, N_j is the number of persons in each score group, X_n is the observed response for person n, and P_n is the predicted response for person v. The unweighted between fit standard deviation can be approximated by

$$S[MS(UB)_i] = \left[\frac{2}{(J-1)}\right]^{1/2}$$ (where J = the number of score groups).

The Logit Residual Index (LRI) is a measure of how far an item characteristic curve (ICC) for an individual item deviates (flatness or steepness) from the common ICC fitted for all items (Mead, 1976). It also indicates the linear trend in the residuals for each item. It can be calculated as

$$\text{LRI}_i = \frac{\sum_{n=1}^{N} (Y_{ni} - Y_{.i})(b_n - d_i)}{\sum_{n-1}^{N} (b_n - d_i)^2}$$

where d_i is the difficulty of the item, b_n is the ability of the person, and N is the number of persons; where Y_{ni} is the standardized residual calculated as:

$$Y_{ni} = \frac{(X_{ni} - P_{ni})}{P_{ni}(1-P_{ni})},$$

with X_{ni} the observed response and P_{ni} the predicted response.

Items with LRI values greater than zero will have an item characteristic curve (ICC) that is steeper than the modeled curve, and items with values less than zero will have an ICC that is flatter than the modeled curve. Negative LRI values indicate that low-ability groups should have positive residuals and high-ability groups should have negative residuals. This indicates that low-ability persons performed better than expected and high-ability persons performed less well than expected. Positive LRI values indicate that low-ability groups should have negative residuals and high-ability groups should have positive residuals. That is, high-ability persons performed better than expected and low-ability persons performed less well than expected. Items with negative total fit statistics tend to have steeper observed ICCs than predicted, indicating an overfit to the model, and items with positive total fit statistics tend to have flatter observed ICCs than predicted, indicating an underfit to the model.

Methods

Simulated Data Sets

Data sets were generated using SIMTEST version 2.1, a software program developed by Luppescu (1992) for simulating dichotomous test data. Six tests were simulated, three tests with normally distributed item difficulties and three tests with uniformly distributed item difficulties. The parameters used to simulate the normally distributed data sets were (a) a mean person ability of 1; (b) a standard deviation of 2; (c) a slope of 1; (d) 100 items; (e) 100 persons, and (f) three levels of guessing [no guessing (0%), 25%, and 50%]. For the uniformly distributed data sets, the parameters were (a) a mean person ability of 1; (b) a standard deviation of 1; (c) a slope of 1, (d) 100 items, (e) 100 persons, and (f) three levels of guessing [no guessing (0%), 25%, and 50%]. A BIGSTEPS (Wright and Linacre, 1992) control program was written to read the dichotomous data and output a data file containing item and person parameters to be read by IPARM (Smith, 1991a) which output a data file containing Rasch item fit statistics and LRI values.

Analysis

An analysis of variance (ANOVA) followed by a Scheffe mean pairwise comparison at the .05 level of significance were used to determine if guessing had an effect on the mean item scores and mean item fit

statistics by levels of guessing (0%,25% and 50%) for each distribution type. To determine if the levels of guessing had an effect on the mean item scores for the distribution types (normal and uniform), mean pairwise comparisons were made between the mean item scores at the same level of guessing. To determine if the level of guessing had an effect on the mean item fit statistic for each distribution type, mean pairwise comparisons were made between the mean item fit statistics at the same level of guessing. An observation of the change (positive /negative) in LRI values associated with misfitting items detected by each item fit statistic (weighted total, unweighted total, and unweighted between) was used to determine the effects of guessing on the LRI.

Results

Mean item score and item fit statistics by levels of guessing for normally distributed item difficulties are in Table 1. Mean item score and item fit statistics by levels of guessing for uniformly distributed item difficulties are in Table 2.

Table 1
Mean Item Score and Item Fit Statistics by Levels of Guessing for Normally Distributed Item Difficulties

				Item Fit Statistics					
		Item Score		UnWt. Total		Wt. total		Between	
Test	% Guessing	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
1	0	.63	.27	.07	.75	.05	.60	.01	.95
2	25	.64	.28	.08	.99	-.02	.86	.22	.97
3	50	.63	.29	.08	.69	.01	.58	.00	.83

Table 2
Mean Item Score and Item Fit Statistics by Levels of Guessing for Uniformly Distributed Item Difficulties

				Item Fit Statistics					
		Item Score		UnWt. Total		Wt. total		Between	
Test	% Guessing	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
4	0	.68	.11	.07	.86	-.02	.76	-.04	.80
5	25	.67	.11	.00	.84	.02	.90	.08	.90
6	50	.73	.10	.02	.75	.03	.59	.04	.86

No significant differences were found at the .05 level between the mean item scores or the mean item fit statistics (unweighted total, weighted total, and unweighted between fit statistics) by the levels of guessing for tests constructed with normally distributed item difficulties (F=.003, df=2,297 - item score; F=.010; df=2,297 - unweighted total; F=.186, df=2,297 - weighted total; and F=1.80, df=2,297 - unweighted between fit, respectively). A significant difference at the .05 level was observed between mean item scores for tests constructed with uniformly distributed item difficulties (F=8.45, df=2,297), but not for the item fit statistics (F=.162, df=2,297 - unweighted total; F=.149, df=2,297 - weighted total; F=.508, df=2,297 - unweighted between fit, respectively). A Scheffe mean pairwise comparison indicated that the mean item scores at the 50% level of guessing were significantly different from those observed at the 0% and 25% levels (Table 3).

Table 3
Scheffe pairwise comparisons for mean item scores (uniform distribution)

Level/Mean	N	S.D.	df	*t*	sig.
0% = .68	100	11	198	.773	.440
25% = .67	100	11			
0% = .68	100	11	198	-3.160	.002
50% = .73	100	10			
25% = .67	100	11	198	-3.966	.000
50% = .73	100	10			

Misfitting Items and the Logit Residual Index (LRI)

A summary of misfitting items and associated fit statistics is presented in Table 4. Seven items (two at 0%, and five at 25%) were detected as misfitting on tests constructed with normally distributed item difficulties. The logit item difficulties ranged from -4.37 (very easy) to 1.46 (fairly difficult). For the tests constructed with uniformly distributed item difficulties, 12 items were detected as misfitting (4 at 4%, 5 at 25% and 3 at 50%). The logit item difficulties for these items ranged from -1.15 to .96. The majority of misfitting items were detected by the unweighted total fit statistic.

Table 4
Summary of Misfitting Item Statistics for each Test

Item #	Logit item diff.	Point. Bis. Corr.	Unwt. Total fit	Wt. Total fit	Abilitly between fit	Mean item score	Logit Residual Index
Test 1							
3	-4.37	-0.09	1.98	0.38	2.70	0.99	-0.26
36	-0.80	0.36	0.94	0.32	2.14	0.80	-0.10
Test 2							
51	0.15	0.27	2.47	1.33	1.36	0.64	-0.67
66	0.41	0.23	2.64	2.20	1.28	0.59	-0.76
72	1.36	0.38	0.40	0.84	2.09	0.40	-0.06
75	1.36	0.60	-2.03	-2.01	-2.01	0.40	0.43
84	1.46	0.66	-2.72	-2.96	2.60	0.38	0.50
Test 3							
None							
Test 4							
11	-1.15	0.26	2.22	0.53	0.66	0.85	-0.41
52	0.01	0.26	1.40	1.97	2.36	0.69	-0.27
67	0.34	0.29	3.21	1.50	0.99	0.63	-1.06
76	0.23	0.39	2.20	0.43	-0.04	0.65	-0.77
Test 5							
49	0.25	0.29	1.54	1.66	2.21	0.63	-0.34
76	0.76	0.26	1.62	2.49	1.92	0.53	-0.36
86	0.91	0.31	2.09	1.49	-0.08	0.50	-0.64
90	0.86	0.29	2.13	1.70	0.26	0.51	-0.67
99	0.96	0.26	1.73	2.40	2.40	0.49	-0.42
Test 6							
16	-0.63	0.20	2.16	0.56	0.06	0.83	-0.36
86	0.76	0.26	1.05	1.74	2.78	0.60	-0.20
89	0.76	0.36	2.10	-0.01	-0.19	0.60	-0.75

Note: Test 1 (0 % guessing, normally distributed item difficulties)
Test 2 (25% guessing, normally distributed item difficulties)
Test 3 (50% guessing, normally distributed item difficulties)
Test 4 (0% guessing, uniformly distributed item difficulties)
Test 5 (25% guessing, uniformly distributed item difficulties)
Test 6 (50% guessing, uniformly distributed item difficulties)

Guessing had an indirect effect on the LRI. The LRI statistic was sensitive to the high positive misfit values associated with the unweighted total fit statistic than to similar values associated with the other item fit

statistics. The greatest magnitude of change was observed when the unweighted total fit statistic misfit value was highly positive (> 2.4) producing a negative LRI value (flatter ICCs than model curve). This indicates that low-ability persons performed better than expected and high-ability persons performed less well. Since only two ability groups were used in the analysis, interpretation of the results were made relatively easy. All fit statistics with positive misfit values produced negative LRI values and those with negative misfit values produced positive LRI values.

Conclusions

As the levels of guessing increased, no significant differences were found between mean item fit statistics (unweighted total, weighted total, and unweighted between fit statistics) for each distribution type (normal and uniform). The mean item scores by levels of guessing were significantly different at the .05 level for tests constructed with uniformly distributed item difficulties, but not normally distributed item difficulties.

It was hypothesized that as the level of guessing increased, the mean item score would increase. This hypothesis held true for tests with uniformly distributed item difficulty distributions, but was not different for tests constructed with normally distributed item difficulties. This finding indicated that tests constructed with a wide range of item difficulties (i.e., normal distribution) tend to stabilize the effects of guessing on the mean item score. However, for tests constructed with a narrow range of difficulties (i.e., uniform distribution), relatively high ability persons tended to consistently guess the correct answer as the probability level increased. This was especially evident when the logit item difficulties were centered on the mean logit ability.

The unweighted total fit statistic was more sensitive to item misfit problems farther away from the mean ability used to simulate the data. These items tended to be either very easy or very difficult. The weighted total fit statistic was more sensitive to item misfit problems centered on the mean ability used to simulate the data. The number of items detected by the between fit statistic increased as the probability of guessing increased. The between fit statistic was therefore more sensitive to item familiarity bias induced by the increased probability of guessing the correct answer. Detected items were functioning quite differently among the two ability groups (high ability and low ability) used in the analyses. These

findings indicated that the between fit statistic when applied to a teacher made test, can be a quick and useful tool for item analysis and test construction.

The LRI was sensitive to changes in the item misfit value associated with the unweighted total fit statistic. High positive item misfit values associated with the unweighted total fit statistic produced a greater change (negative) in the LRI value than similar item misfit values associated with the other Rasch item fit statistics. The usefulness of the LRI lies in the fact that it can identify the linear trend in the residuals for each item by demographic/biographical characteristics. This permits quick identification of differential item functioning among subgroups. Therefore, in an effort to obtain an accurate measurement of individual differences, the LRI appears to be an essential tool for item analysis and test construction.

References

Andersen, E. B. (1973). Goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140.

Andrich, D. (1988). *Rasch models for measurement.* Sage University paper series on Quantitative Applications in the Social Sciences (Series No. 07-068). Beverly Hills: Sage.

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory.* New York: Wiley.

Ferguson, G. A. (1981). *Statistical analysis in psychology and education.* New York: McGraw-Hill.

Glaser, R. (1949). A methodological analysis of the inconsistency of responses to test items. *Educational and Psychological Measurement, 9,* 721-739.

Glaser, R. (1952). The reliability of inconsistency. *Educational and Psychological Measurement, 11,* 60-64.

Gustafsson, J-E. (1980). Testing and obtaining fit of data to the Rasch model. *The British Journal of Mathematical and Statistical Psychology, 33,* 205-233.

Habermann, S. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics, 5,* 815-841.

Luppescu, S. (1992). SIMTEST 2.1 [Computer program for simulating test data]. Chicago.

Mead, R.J. (1976). Assessing the fit of data to the Rasch model through the analysis of residuals. Unpublished Ph.D. Dissertation, University of Chicago.

Mosier, C.I. (1941). Psychophysics and mental test theory: II. The constant process. *Psychological Review, 47,* 235-249.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical statistics and probability*. Berkeley: University of California Press, *4*, 321-333.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Rev. ed.). Chicago: University of Chicago Press.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 45*, 433-444.

Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement, 46,* 359-372.

Smith, R. M. (1988a). *A comparison of the power of Rasch total and between item fit statistics to detect measurement disturbances*. A paper presented at the annual meeting of the American Educational Research Association, New Orleans.

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

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

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

Smith, R. M., and Hedges, L.V. (1982). A comparison of the likelihood ratio χ^2 and Pearsonian c^2 tests of fit in the Rasch model. *Educational Research and Perspectives, 9,* 44-54.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques.* New York: Teachers College Press.

Thurstone, L. L., and Chave, E. J. (1929). *Measurement of attitudes.* Chicago: University of Chicago Press.

Wollenberg, A. L. van den. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123-140.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97-116.

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

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

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

JOURNAL OF OUTCOME MEASUREMENT®, 2(4), 351 - 364

Corrected Rasch Asymptotic Standard Errors for Person Ability Estimates

Richard M. Smith
Rehabilitation Foundation, Inc.

Most calibration programs designed for the family of Rasch psychometric models report the asymptotic standard errors for person and item measure estimates resulting from the calibration process. Although these estimates are theoretically correct, they may be influenced by any number of factors, e.g., restrictions due to the loss of degrees of freedom in estimation, targeting of the instrument, i.e., the degree of offset between mean item difficulty and mean person ability, and the presence of misfit in the data. The effect of these factors on the standard errors reported for the person has not been previously reported. The purpose of this study was to investigate the effects of these three factors on the asymptotic standard errors for person measures using simulated data. The results indicate that asymptotic errors systematically underestimate the observed standard deviation of ability in simulated data, though this underestimation is usually small for targeted instruments with reasonable sample size. However, the underestimation can easily be corrected with a simple linear function. These simulations use only dichotomous data and the results may not generalize to the rating scale and partial credit models.

An earlier version of this paper was presented at the Annual Meeting of the National Council on Measurement in Education, New York, April, 1996.

Requests for reprints should be sent to Richard M. Smith, RFI, P.O. Box 675, Wheaton, IL 60187.

Most calibration programs designed for the family of Rasch psychometric models report the asymptotic standard errors for person and item estimates resulting from the calibration process. Although these estimates may be theoretically correct, they can be influenced by any number of factors, such as restrictions due to the loss of degrees of freedom in the estimation process (sample size), offset between the person and item measures, and the presence of misfit in the data. These influences have not been previously investigated. There are two sets of standard errors reported, those for item difficulty and those for person ability estimates. This study is limited to standard errors associated with the estimation of person ability in the dichotomous model.

The general asymptotic standard error (*SE*) for a person with ability *b* that corresponds to raw score *r* in the general Rasch model (Wright and Masters, 1982) is given by:

$$SE(b_r) = \left[\sum_{i}^{L} \left(\sum_{k=1}^{m_i} k^2 p_{rik} - \left(\sum_{k=1}^{m_i} k p_{rik} \right)^2 \right) \right]^{-1/2} \quad (1)$$

where L is the number of items, m_i is the number of steps in item i, and p_{rik} is the probability of a person with a score r responding in category k to item i. In the dichotomous model (Wright and Stone, 1979) this simplifies to:

$$SE(b_r) = \left[\sum_{i}^{L} p_{ri}(1-p_{ri}) \right]^{-1/2} \quad (2)$$

where p_{ri} is the probability that a person with raw score r answers item i correctly.

One of the most frequently used Rasch calibration programs, BIGSTEPS (Wright and Linacre, 1992) contains a command that allows the user to modify the asymptotic standard errors based on the amount of misfit present in that person's response pattern. This correction increases the standard error for persons whose INFIT mean square departs from the expected value of 1.0 The rules for this correction and a computational example are shown in Table 1.

Table 1

BIGSTEPS Standard Error Adjustments

Using REALSE=Y

If INFIT Mean Square >1.0 then SE_r = (INFIT MNSQ)^{½} (SE)

If INFIT Mean Square <1.0 then SE_r = (INFIT MNSQ)$^{-½}$ (SE)

For example, if INFIT MNSQ = 1.41 and SE = 0.59 then SE_r = $(1.41)^{½}(0.59) = .70$

The FACETS program (Linacre and Wright, 1993) also contains a control statement that can be used to select either asymptotic or enlarged standard errors. The term enlarged refers to the use of the person or item mean square to adjust asymptotic standard error for the presence of misfit in the response data.

The purpose of this study is to examine empirically, through the use of simulated data that fit the dichotomous Rasch model, the relationship between the reported asymptotic standard errors for person ability measures and the observed standard deviation of estimated ability of simulated persons with the same generating ability. The magnitude of the systematic difference between the asymptotic values and the observed values and possible corrections are reported. Finally, the effect of misfitting data caused by guessing when the correct answer is not known, on the observed standard deviations of the person ability estimates were examined and the appropriateness of the INFIT correction currently used in BIGSTEPS and FACETS was evaluated.

Methods

Simulated data generated to fit the dichotomous Rasch model were used in this study to determine the empirical standard deviation of estimated person ability in samples with identical generating ability within each sample. It may seem restricting to limit each simulation to a single ability, but the nature of the design, comparing the mean asymptotic standard error to the observed standard deviation of the estimated ability, requires this type of restriction. The effect on different abilities was investigated by varying the generating person ability in different samples. The simulations were limited to a five test lengths (five, ten, twenty, forty, and eighty item tests). (See Table 2 for a description of all of the simu-

lated data sets.) One thousand persons were used in each simulation. The generating abilities used in the study were limited to nine values that ranged from -2.5 logits to 1.5 logits in 0.5 logit steps. The item difficulty distribution used in each simulation was uniform and covered the range from -2.0 to +2.0 logits, in 0.2 logit steps, except in the case were there were 10 or

Table 2
Simulation Summary Information

Number of Items 5, 10, 20, 40, 80
Dichotomous Model
1000 simulated persons per ability
Uniform Distribution of item difficulty (-2.0 to +2.0, shown below for 20 items)
Ten replications of each condition
Eight fitting data conditions (-2.0 to +1.5 logit generating abilities) for all test lengths
Seven guessing data conditions (-2.5 to +0.5 logit generating abilities) only for a 20 item test
Six combined (fitting and guessing) data sets (n=2000, -2.0 to +0.5 logit generating abilities) only for a 20 item test

Possible Values:	
Item Difficulty	Person Ability
2.0	-2.5
1.8	-2.0
1.6	-1.5
1.4	-1.0
1.2	-0.5
1.0	0.0
0.8	0.5
0.6	1.0
0.4	1.5
0.2	
-0.2	
-0.4	
-0.6	
-0.8	
-1.0	
-1.2	
-1.4	
-1.6	
-1.8	
-2.0	

fewer items. For ten or fewer items, the items were uniformly spread across the -2.0 to +2.0 range in larger steps. The results reported represent the average of ten replications for each combination of generating ability and test length. Results are reported for eight levels of person ability across the five test lengths when the data fit the model. In the guessing simulations seven levels of person ability were used with a single test length, twenty items.

In the guessing simulations, guessing was introduced by setting the probability of a correct response equal to 0.25 (the probability of randomly guessing correctly on a four-choice item) whenever the modeled probability was lower than 0.25. Depending on the person ability used in the simulations this would result in guessing on between 3 and 17 of the

Table 3
Guessing When Not Known (If p<.25 then p=.25)

Item		Person Ability						
Diff.	1.0	0.5	0.0	-.5	-1	-1.5	-2.0	-2.5
2.0	.	x	x	x	x	x	x	x
1.8	.	x	x	x	x	x	x	x
1.6	.	x	x	x	x	x	x	x
1.4	.	.	x	x	x	x	x	x
1.2	.	.	x	x	x	x	x	x
1.0	.	.	.	x	x	x	x	x
0.8	.	.	.	x	x	x	x	x
0.6	.	.	.	x	x	x	x	x
0.4	.	.	.	.	x	x	x	x
0.2	.	.	.	.	x	x	x	x
-0.2	.	.	.	.	.	x	x	x
-0.4	.	.	.	.	.	x	x	x
-0.6	.	.	.	.	.	.	x	x
-0.8	.	.	.	.	.	.	x	x
-1.0	.	.	.	.	.	.	.	x
-1.2	.	.	.	.	.	.	.	x
-1.4	.	.	.	.	.	.	.	x
-1.6	.	.	.	.	.	.	.	.
-1.8	.	.	.	.	.	.	.	.
-2.0	.	.	.	.	.	.	.	.
Guessing Items	0	3	5	8	10	12	14	17

20 items. For example, if the persons were at a generating ability of 0.0 logits, there would be guessing on the five items with a difficulty higher than 1.2 logits. For persons with a generating ability of -2.0 there would be guessing on the 14 items with a difficulty of -0.8 logits or higher. The item difficulties with possible distortion due to guessing are listed in Table 3 for the seven guessing data sets.

The simulated responses were calibrated with the BIGSTEPS program. The mean and standard deviation of the estimated standard errors and the mean and standard deviation of the estimated abilities were then compared to determine if the asymptotic standard errors reported by the program were accurate. The *SD/SE* ratio was used to evaluate the agreement between the mean asymptotic standard errors reported for the estimated person measures and the observed standard deviation of the estimated person ability measures. In this ratio *SD* represents the standard deviation of the estimated measures for the 1000 persons simulated to have the same ability. *SE* represents the mean asymptotic standard error reported for these estimated measures. Values greater than 1.00 indicate that, on the average, the observed standard deviation of the estimated measures for the 1000 persons was greater than the average asymptotic standard error reported for the estimated measures of those persons. A value greater than 1.00 suggests that the asymptotic standard errors underestimates the dispersion of the persons due to the errors of estimation.

Results

For the simulations with five item tests over five different person generating abilities (-1.0 to +1.0) , the average *SD/SE* ratio was 1.093. The full range of person abilities was not used in these simulations due to the frequency of zero and perfect scores for the extreme generating person abilities. For the simulations with ten item tests over seven different generating abilities (-1.5 to +1.5), the average *SD/SE* ratio was 1.054. For the simulations with twenty item tests over eight generating abilities (-2.0 to +1.5), the average *SD/SE* ratio was 1.037. One additional person ability was added to match the guessing simulations reported later. For the simulations with forty item tests over seven generating abilities (-1.5 to +1.5), the average *SD/SE* ratio was 1.011. For the simulations with eighty item tests over seven generating abilities (-1.5 to +1.5), the average *SD/SE* ratio was 1.007. A graph showing the *SD/SE* ratio for all of the values reported above is found in Figure 1. A pair of reference lines are shown at 1.02 and 0.98 to provide a frame of reference.

There is an obvious progression in the *SD/SE* ratio based on the number of items on the test. The underestimation of the standard error of the ability estimate is closely approximated by the function $(2L/(2L\text{-}1))$, where L is the number of items, which would yield corrections of 1.11 for the five item tests, 1.053 for the ten item tests, 1.026 for the twenty item

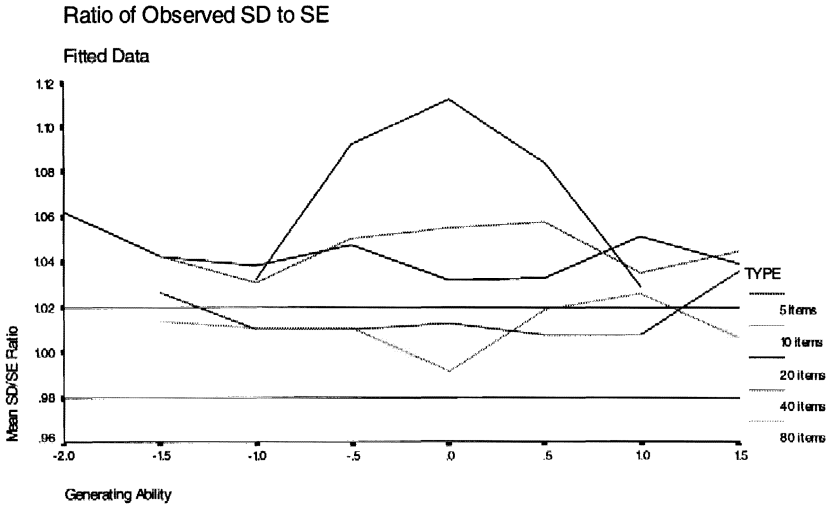


Figure 1. Ratio of observed *SD* to mean *SE* - simulated fitting data.

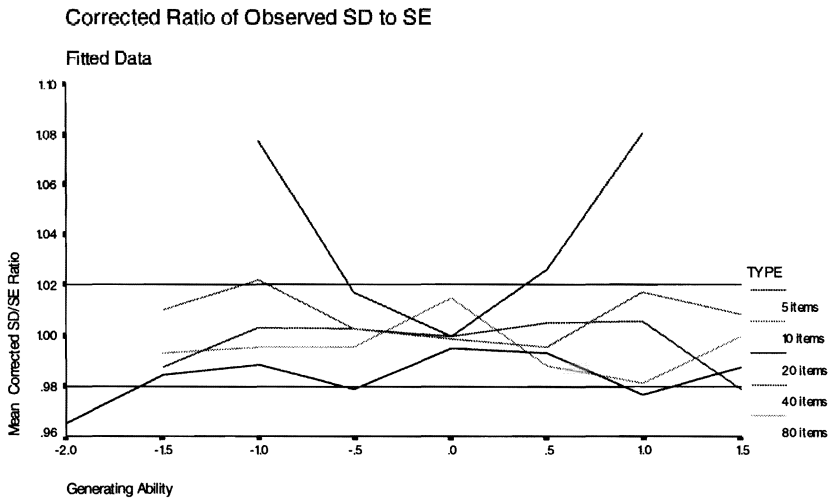


Figure 2. Corrected Ratio of Observed *SD* to *SE*

tests, 1.013 for the forty item tests and 1.0006 for the eighty item tests. Using the corrections ($2L/(2L-1)$) to increase the value of the estimated asymptotic standard errors would yield a *SD/SE* ratio of approximately 1.0, with all of the observed ratios falling in the range of 0.98 to 1.02. The results for the corrected *SD/SE* ratios are shown in Figure 2. The only exceptions are the cases where the number of perfect and zero scores

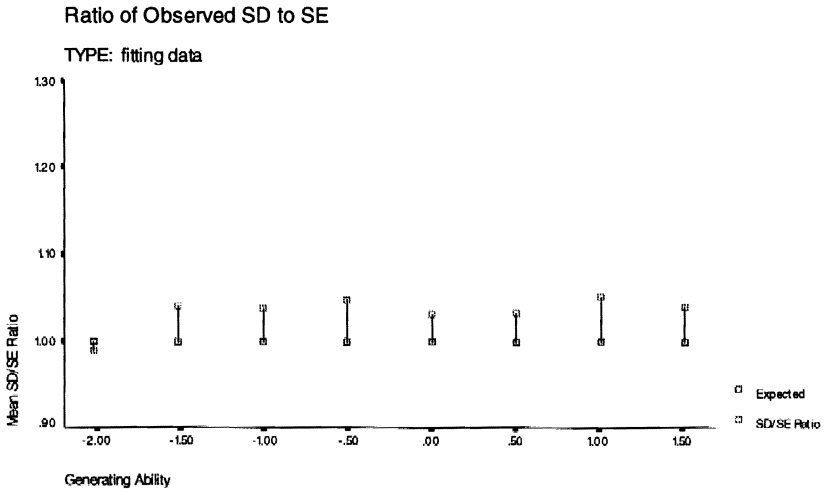


Figure 3. Ratio of observed *SD* to mean *SE* - simulated fitting data

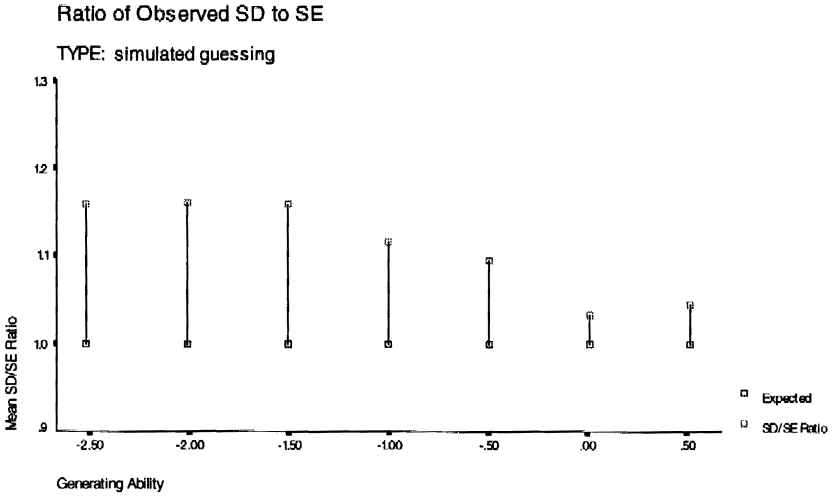


Figure 4. Ratio of observed *SD* to mean *SE* - simulated guessing data

in the data increase, as with the 5 item test, so that the pseudo-ability estimates assigned to the persons with zero or perfect raw scores cause the observed standard deviation to be artificially high. There appears to be no systematic bias introduced by the offset between the mean item difficulty and the generating person ability.

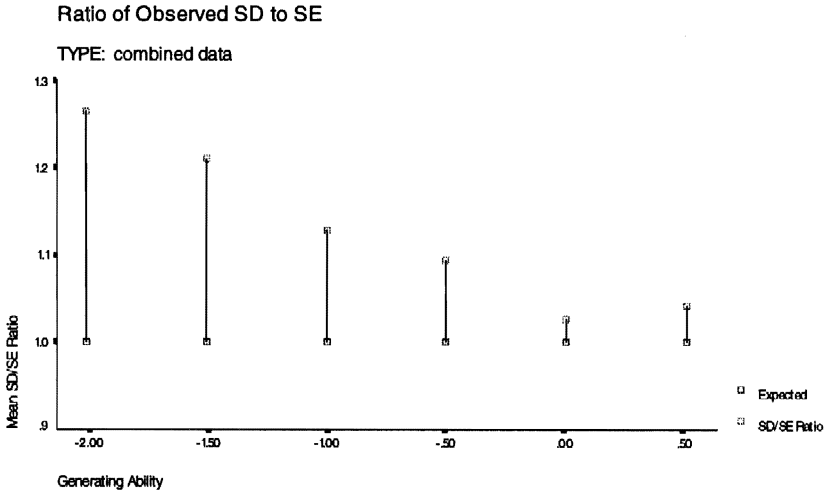


Figure 5. Ratio of observed *SD* to mean *SE* - combined data

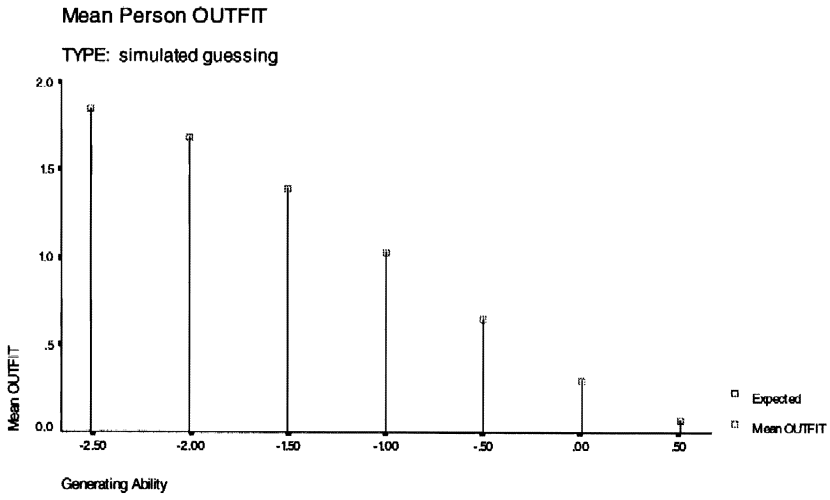


Figure 6. Mean person OUTFIT statistic - simulated guessing data

The results of the guessing simulations indicated that the presence of guessing in the data decreased the mean asymptotic estimates of the standard errors. This is as expected due to the fact that the guessing increased the estimated person measures, moving it closer to the center of the item difficulty distribution and decreasing the asymptotic standard

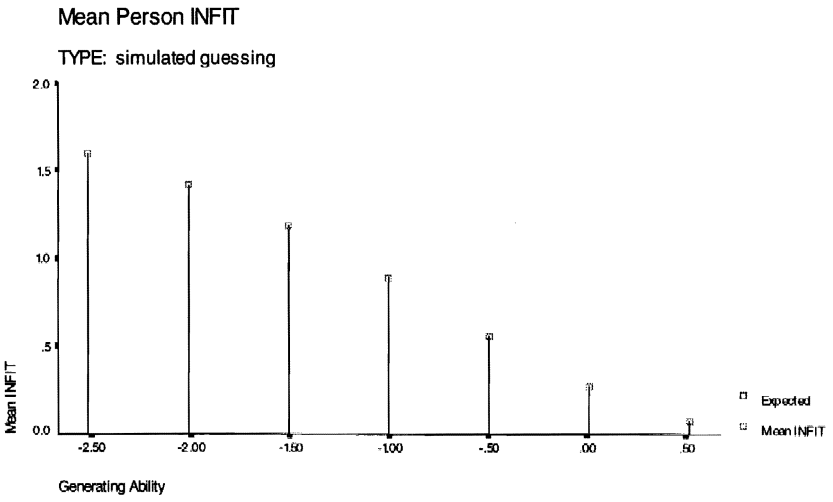


Figure 7. Mean person INFIT statistic -simulated guessing data

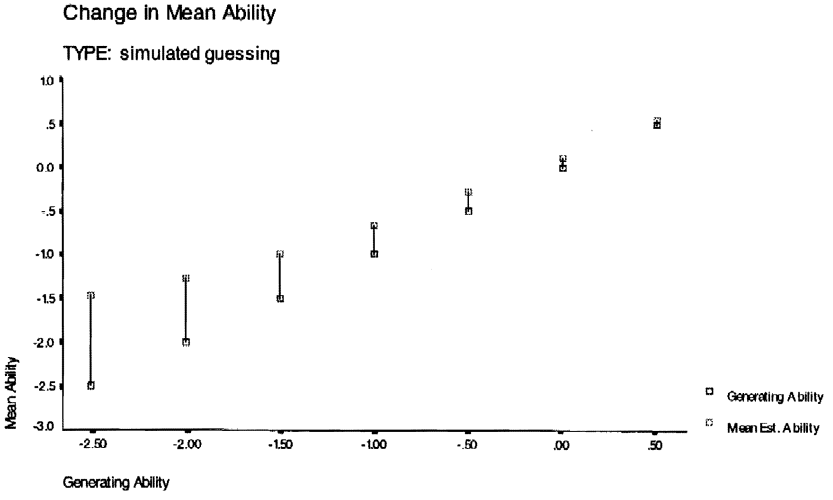


Figure 8. Change in mean person measure - simulated guessing data

error. For the -1.5 logit generating ability simulations, the mean asymptotic *SE* for the 10 fitting simulations was 0.62 while the mean person measure *SD* was 0.644. For the 10 -1.5 logit generating ability simulations with guessing, the mean asymptotic *SE* was 0.57 while the mean person measure *SD* was 0.658. The presence of simulated guessing reduced the mean asymptotic *SE* while slightly increasing the standard de-

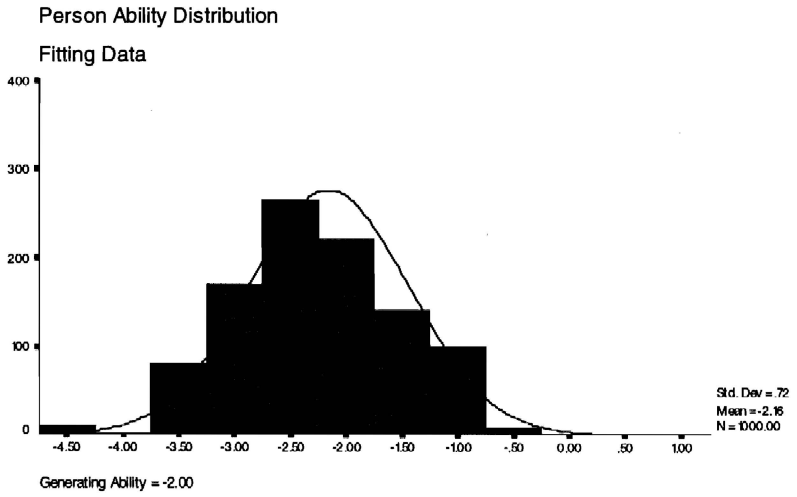


Figure 9. Person ability distribution - fitting data

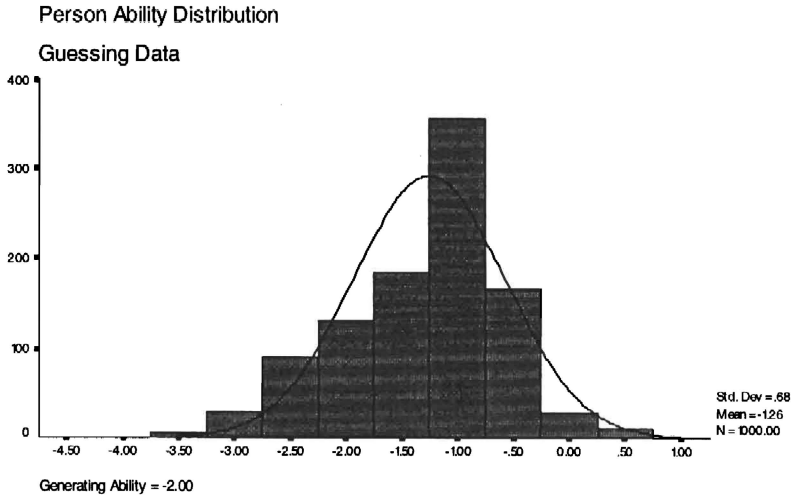


Figure 10. Person ability distribution - simulated guessing data

viation of the estimated person measures. Thus the *SD/SE* ratios were higher for the guessing data sets then for the fitted data sets. To illustrate this point, the mean *SD/SE* ratios for fitting data are shown in Figure 3, while the mean (over 10 replications) *SD/SE* ratios for the guessing data are shown in Figure 4. For the guessing data alone the *SD/SE* ratio varied between 1.17 and 1.03. This was considerably higher that the average *SD/SE* ratio reported for the fitted data on the twenty item test (1.037).

There is a clear relationship between the amount of guessing present in the data and the size of the *SD/SE* ratio as seen in Figure 4. As the proportion of guessing data decreased, moving from a generating ability of -2.5 logits to an ability of 0.5 logits, the size of the *SD/SE* ratios also decreased. If the mean INFIT value for these persons is used to modify the asymptotic standard error as shown in Table 1, the *SD/SE* ratios return to approximately that found with the fitted data sets.

When the fitted data and the guessing data were combined in a single analysis for each ability level to represent the case where the sample had a mixture of guessing and fitted data (n=2000, with 10 replications), there was a further increase in the *SD/SE* ratio. These results are shown in Figure 5. In these analysis the *SD/SE* ratios ranged from 1.25, for the -2.00 generating ability set, down to 1.04, for the 0.5 generating ability set. As in the guessing analysis there was a clear trend suggesting that as the proportion of guessing in the data increased, the *SD/SE* ratio increased. The use of the BIGSTEPS INFIT correction in this case reduced the *SD/SE* ratio, but not back to the level found in the fitting data.

The amount of misfit in the guessing simulations can be seen in Figures 6 and 7. The mean person INFIT (weighted) and OUTFIT (unweighted) person fit statistics (standardized using the cube root transformation of the mean squares) by generating ability. As the generating ability increases and the resulting targeting of the items on the person improves, reducing guessing, there is a lower mean fit statistic. The expected value of this statistic when the data fit the model is 0.00. Interestingly, as the proportion of guessing increases, the mean values for the INFIT and OUTFIT diverge, with the OUTFIT values increasing more rapidly than the INFIT values. This indicates that the OUTFIT statistic is slightly more sensitive to the guessing simulated in the data.

The effect of the presence of guessing on the estimated person abilities can be seen in Figure 8. As the amount of guessing increases, the estimated mean person abilities are increasingly higher then the generat-

ing values. The shift in the distribution results not only in an increased mean of the sample, but a shift in the skewness of the distribution. This difference can be seen by comparing Figures 9 and 10 (These graphs are based on a single fitted and guessing simulation with a generating person measure of -2.0). There was a pronounced negative skew to the guessing data shown in Figure 10. It is interesting to note the underestimate of person ability in the fitting data shown in Figure 9. In this case the generating value was -2.0. While mean estimated person measure was -2.16. This is the result of the number of zero scores, represented by the block of persons with an estimated ability of -4.5, who were arbitrarily assigned a measure since there is no true ability estimate for zero raw scores.

Conclusions

These results apply only to the unconditional estimation procedure used in BIGSTEPS as applied to dichotomous data. The asymptotic standard errors reported by most programs are usually thought of as a lower bound for the standard error. But it seems counter productive to report a number that is known to be smaller than it actually is. If classification decisions are based on a measure ±2 standard errors, then the person standard error used to make a decision should reflect the best estimate for that situation, not the lowest value that standard error value could take. It can be argued that this might make a small difference, but the change of a single raw score point on a high stakes examination can be very important to the examinees who fail using one standard error and pass using another.

These results highlight two problems. First, the asymptotic standard errors currently reported in BIGSTEPS underestimate the true variation in persons generated to have the same ability by a factor of $(2L/(2L\text{-}1))$. This bias may be the result of the UCON bias correction $(L\text{-}1)/L$ option available in BIGSTEPS (Wright and Douglas, 1997). This correction was used in the estimation of the item difficulties from the simulated data. With longer tests this underestimation may not be of practical importance, but with shorter tests it may pose problems. A correction such as this could easily be built into the estimation program. Second, the presence of guessing in the response patterns causes further underestimation of the asymptotic standard errors. This underestimation is further increased when there is a mixture of fitting and misfitting response patterns. It seems reasonable to use a correction for the asymptotic standard

error based on one of the person fit indices that are calculated after the calibration. The choice of fit indices to correct the estimates of the asymptotic standard errors is important. There is a difference in the mean INFIT and OUTFIT statistics for the simulated guessing data. The differences increase markedly as the amount of guessing increases. The mean outfit for the -2.00 sample was 1.95, whereas the INFIT mean for the same sample was 1.55. For the 0.50 logit ability sample the means are approximately equal. The use of the person INFIT or OUTFIT statistic for the asymptotic standard error correction in BIGSTEPS can result in different corrections depending upon the amount of guessing present in the response pattern. These simulations results suggest that a correction for misfit based on the OUTFIT statistic would yield corrected asymptotic standard errors that are closer to the standard deviation of simulated cases with the same generating person ability.

Acknowledgment

The author would like to thank Benjamin D. Wright and two anonymous reviewers for their suggestions which improved the manuscript.

References

Andrich, D. (1988). *Rasch models for measurement.* Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-068. Beverly Hills: Sage Publications.

Linacre, J.M. and Wright, B.D. (1993). *A user's guide to FACETS: Many-facet Rasch analysis.* Chicago: MESA Press.

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

Wright, B.D. and Linacre J.M. (1992). *A user's guide to BIGSTEPS: A Rasch model computer program.* Chicago: MESA Press.

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

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

Journal of Outcome Measurement®
Volume 2
Author and Title Index

CONTRIBUTOR INFORMATION

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

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

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

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

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

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

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

SUBSCRIBER INFORMATION

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

 ISSN 1090-655X.