Designation: E 2171 – 02

An American National Standard

Standard Practice for Rating-Scale Measures Relevant to the Electronic Health Record[1]

This standard is issued under the fixed designation E 2171; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 This standard addresses the identification of data elements from the EHR definitions in Guide E 1384 that have ordinal scale value sets and which can be further defined to have scale-free measurement properties. It is applicable to data recorded for the Electronic Health Record and its paper counterparts. It is also applicable to abstracted data from the patient record that originates from these same data elements. It is applicable to identifying the location within the EHR where the observed measurements shall be stored and what is the meaning of the stored data. It does not address either the uses or the interpretations of the stored measurements.

2. Referenced Documents

2.1 *ASTM Standards:*

E 177 Practice for Use of the Terms Precision and Bias in ASTM Test Methods[2]

E 456 Terminology Related to Quality and Statistics[2]

E 691 Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method[2]

E 1169 Guide for Conducting Ruggedness Tests[2]

E 1384 Guide for Content and Structure of the Computer-Based Patient Record[3]

3. Terminology

3.1 *Definitions*—Full definitions and discussion of Scale-Free Measurement Terms are given in Annex A1.

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *adaptive measurement*—advantage of measurement to account for missing data.

3.2.2 *additivity*—rating scale adherence to associativity and commutability.

3.2.3 *bias analysis*—investigation of considerations relative to subject or area of performance.

3.2.4 *calibration*—process of establishing additivity and reproducability of a data set.

3.2.5 *concatenation*—process of measurement uses enumerated physical unit quantities equal to the magnitude of the measured item.

3.2.6 *construct*—name of the conceptual domain measured.

3.2.7 *convergence*—closing of the differences in sequential measure estimates.

3.2.8 *counting*—basic activity upon which measurement is based and utilizes enumeration.

3.2.9 *data*—observation made in such a way that they lead to generalization.

3.2.10 *data quality/ statistical consistency/ model fit*—establishment of whether the measuring instrument is affected by the object of measurement.

3.2.11 *determinism*—measurement model that requires counts to be sufficient for reproducing the pattern of the responses over the length of the instrument.

3.2.12 *dimensionality*—property of having multiple components of a measured value.

3.2.13 *equality/cocalibration*—process of ensuring that different instruments measure the same property.

3.2.14 *error*—uncertainty of measured properties.

3.2.15 *estimation algorithms*—mathematical specification of an observational framework.

3.2.16 *incommensurable/commensurable*—measure value of the same quantity does/does not depend upon rating/responses of the rating construct and does not/does remain constant.

3.2.17 *instrument*—sensing device having a defined scale.

3.2.18 *intra and inter-laboratory testing*—variability testing using the same setting/measure/operator as opposed to different setting/measure/operators.

3.2.19 *item response/latent trait theory*—analytic models that forego prescriptive parameter separation, sufficiency and scale and sample free data standards for additional descriptive parameters.

3.2.20 *items/item-bank*—part of survey statements/test questions for adaptive administration.

3.2.21 *levels of measurement*—nature of scale of measurement.

3.2.22 *logit*—scale unit using logarithms of odds ratios. ($P/1-P$).

---

[1] This practice is under the jurisdiction of ASTM Committee E31 on Healthcare Informatics and is the direct responsibility of Subcommittee E31.19 on Electronic Health Record Content and Structure.

Current edition approved Dec. 10, 2002. Published February 2003.

[2] *Annual Book of ASTM Standards*, Vol 14.02.

[3] *Annual Book of ASTM Standards*, Vol 14.01.

3.2.23 *mathematical entities*—concepts that can be taught or learned through what is already known.

3.2.24 *measurement*—determining in units the value of a property in a scale having magnitude (that is, ratio or difference).

3.2.25 *metaphor in measurement*—suspension of disbelief of some areas or properties in the name of estimating magnitude.

3.2.26 *metric*—measure of a property in defined units.

3.2.27 *missing data*—use of uncalibrated data in instruments with varying numbers of items.

3.2.28 *multi-faceted measurement*—use of measurement models that have more than two basic parameters.

3.2.29 *ordinal data*—one scale for measurement.

3.2.30 *population*—universe of elements relevant to measurement of a particular construct.

3.2.31 *probabilistic conjoint measurement*—framework for demonstrating data quality, statistical consistency, and model fit of non-deterministic measures with a stable order of facets.

3.2.32 *quantification*—cocalibration of different constructs with respect to the same property (variable) in a common metric.

3.2.33 *Rasch analysis measurement and models*—analytic model specifying the observational framework and data quality measures for quantification.

3.2.34 *raw score*—sum of ratings or count of direct responses in a given measurement event.

3.2.35 *reliability*—ratio of variation to error or signal to noise.

3.2.36 *repeatability*—variability of measurements in a single setting by a single operator using the same measuring instrument.

3.2.37 *reproducibility*—variability of measurements in different settings.

3.2.38 *root mean square error*—mathematical algorithm for determining the variation due to error of the estimates.

3.2.39 *sample*—subset of measured population.

3.2.40 *sample size*—magnitude of the measured population.

3.2.41 *scale-free/scale-dependent*—measures not affected by the instrument employed as opposed to measures that are so affected.

3.2.42 *separability theorem/parameter separation*—ability of measures to be independent of the instrument selected and ability of the instrument's item calibrations to be independent of the sample measured.

3.2.43 *software*—packages of machine code used for data analysis.

3.2.44 *specific objectivity*—data satisfying the separability theorem.

3.2.45 *standardized*—common conventions for instruments, reference measurement material, scales and units of measure for a measurement process.

3.2.46 *sufficiency*—statistics that extract all available information from the data.

3.2.47 *targeting*—lack of floor an/or ceiling effects in measurement.

3.2.48 *transparency*—ability to "look through" raw scores to the composite ratings producing that score (see also *sufficiency*).

3.2.49 *unit of measurement*—common conventions for the appropriate smallest basic measures for a given construct.

3.2.50 *validity/construct/content*—both content and construct must make sound theoretical sense to be considered valid.

3.2.51 *variable*—attribute of the property being measured.

4. Significance and Use

4.1 The simplicity and practicality of Rasch's probabilistic scale-free measurement models have brought within reach universal metrics for educational and psychological tests, and for rating scale-based instruments in general. There are at least 3 implications to the application of Rasch's models to the health-related calibration of universal metrics for each of the variables relevant to the Electronic Health Record (EHR) that are typically measured using rating scale instruments.

4.1.1 First, establishing a single metric standard with a defined range and unit will arrest the burgeoning proliferation of new scale-dependent metrics.

4.1.2 Second, the communication of the information pertaining to patient status represented by these measures (physical, cognitive, and psychosocial health status, quality of life, satisfaction with services, etc.) will be simplified.

4.1.3 Third, common standards of data quality will be used to evaluate and improve instrument performance. The vast majority of test and survey data quality is currently almost completely unknown, and when quality is evaluated, it is via many different methods that are often insufficient to the task, misapplied, misinterpreted, or even contradictory in their aims.

4.1.4 Fourth, currently unavailable economic benefits will accrue from the implementation of measurement methods based on quality-assessed data and widely accepted reference standard metrics. The potential magnitude of these benefits can be seen in an assessment of 12 different metrological improvement studies conducted by the National Science and Technology Council (Subcommittee on Research, 1996). The average return on investment associated with these twelve studies was 147 %. Is there any reason to suppose that similar instrument improvement efforts in the psychosocial sciences will result in markedly lower returns?

4.2 Until now, it has been assumed that the Guide E 1384 would necessarily have to stipulate fields for the EHR that would contain summary scores from commonly used functional assessment, health status, quality of life, and satisfaction instruments. This is because standards for rating scale instruments to date have been entirely content-based. Those who have sought "gold" or criterion standards that would command universal respect and relevance have been stymied by the impossibility of identifying content (survey questions and rating categories) capable of satisfying all users' needs. Communication of patient statistics between managers and clinicians, or payors and providers, may require one kind of information; between providers and referral sources, other kinds; between providers and accreditors, yet another; among clinicians themselves, still another; and even more kinds of information may be required for research applications.

4.2.1 For instance, payors may want to know outcome information that tells them what percentage of patients discharged can function independently at home. A hospital manager, referral source, or accreditor might want to know more detail, such as percentages of patients discharged who can dress, bathe, walk, and eat independently. Clinicians will want to know still more detail about amounts of independence, such as whether there are safety issues, needs for assistive devices, or specific areas in which functionality could be improved. Researchers may seek even more detail yet, as they evaluate differences in outcomes across treatment programs, diagnostic groups, facilities, levels of care, etc.

4.2.1.1 Members of each of these groups have, at some time, felt that their particular information needs have not been met by the tools designed and developed by members of another group. Despite the fact that the information provided by these different tools appears in many different forms and at different levels of detail, to the extent that they can be shown to measure the same thing, they can do so in the same metric. This is the primary result of the introduction of Rasch's probabilistic scale-free measurement models. The different purposes guiding the design of the instruments will still continue to impact the two fundamental statistics associated with every measure: the error and model fit. More general, and also less well-designed instruments, will measure with more error than those that make more detailed and consistent distinctions. Data consistency is the key to scale-free measurement.

4.3 The remainder of this document (*1*) identifies, in Section 5, the fields in the current Guide E 1384 targeted for change from a scale-dependent to a scale-free measurement orientation; (*2*) lists referenced ASTM documents; (*3*) defines scale-free measurement terms, often contrasting them with their scale-dependent counterparts; (*4*) addresses the significance and use of scale-free measures in the context of the EHR; (*5*) lists, in Annex A2, scientific publications documenting relevant instrument calibrations; (*6*) briefly presents some basic operational considerations; (*7*) lists minimum and comprehensive arrays of EHR database fields; and (*8*) lists, in Annex A3, the references made in presentation of the measurement theory, estimation methods, etc.

4.4 Publications of calibration studies referencing this practice and the associated standard practice should require:

4.4.1 The use of measures, not scores, in all capture of data from the EHR for statistical comparisons;

4.4.2 The reporting of both the traditional reliability statistics (Cronbach's alpha or the KR20) and the additive, linear separation statistics (Wright & Masters, 1982), along with their error and variation components, for both the measures and the calibrations;

4.4.3 A qualitative elaboration of the variable defined by the order of the survey questions or test items on the measurement continuum, preferably in association with a figure displaying the variable;

4.4.4 Reporting of means and standard deviations for each of the three essential measurement statistics, the measure, the error, and the model fit;

4.4.5 Statement of the full text of at least a significant sample of the questions included on the instrument;

4.4.6 Specification of the mathematical model employed, with a justification for its use;

4.4.7 Specification of the error estimation and model fit estimation algorithms employed, with mathematical details and justification provided when they differ from those routinely used;

4.4.8 Evaluation of overall model fit, elaborated in a report on the details of one or more of the least and most consistent response patterns observed;

4.4.9 Graphical comparison of at least two calibrations of new instruments from different samples of the same population to establish the invariance of the item calibration order across samples;

4.4.10 Graphical comparison of measures produced by at least two subsets of items on new instruments to establish the invariance of the person measure order across scales (collections of items);

4.4.11 Graphical comparison of new instrument calibrations with the calibrations produced by other instruments intended to measure the same variable in the same population, to establish the potential for sample-free equating of the instruments and establishment of reference standards;

4.4.12 At least a useable prototype of the instrument employed, with the worksheet laid out to produce informative quantitative measures (not summed scores) as soon as it is filled out; and

4.4.13 Graphical presentation of the treatment and control groups' measurement distributions, for the purpose of facilitating a substantive interpretations of differences' significance.

5. Applicable Data Elements

5.1 The data elements in Guide E 1384 which are affected by the suggestions for measurement standardization made here include the following:

PHYSICAL EXAM SEGMENT
09001.16 Patient Health Status Measure Name
09001.17. Patient Health Status Measure Total Value
09001.19. Patient Health Status Measure Element Name (M)
90001.19.01 Patient Health Status Measure Element Value

ENCOUNTER RECEIPT SUBSEGMENT
14001.A154. Patient Receipt Health Status Measure Name
14001.A156. Patient Receipt Health Status Measure Total Value
14001.A160. Patient Receipt Health Status Measure Element Name (M)
14001.A160.01. Patient Receipt Health Status Measure Element Value

ENCOUNTER DISPOSITION SUBSEGMENT
14001.F067. Patient Disposition Health Status Measure Name
14001.F068. Patient Disposition Health Status Measure Total Value
14001.F069. Patient Disposition Health Status Measure Element Name
14001.F069.01. Patient Disposition Health Status Measure Element Value

5.1.1 The reason why the 09001.16.—09001.19.1. data elements are present is to allow capture of one-time uses of the measuring instrument that are not clearly related to an admission or discharge assessment. The encounter receipt and disposition subsegments enable capture of health status at the beginning and end of major encounters. The core health status

rating scale variables that must be addressed include physical function, cognitive function, psychosocial function, quality of life, and satisfaction with services. Each of these five variables will require a bare minimum of two indicators per assessment: a measure and an error. Both of these numbers can be obtained from specially-designed paper-and-pencil worksheets for calibrated instruments (for example, survey forms). As the assessments are increasingly performed with, and recorded directly into, handheld computers, the computer can be used to calculate and store one or more data quality indicators along with the measure and error.

5.1.2 All of these statistics will be stored for each assessment of each variable (data element instance).

5.1.3 Disease-specific measures of functional health status will also be required. There are many such survey measurement instruments in circulation but few have yet been studied using scale-free methods. Diseases for which such instruments exist include alcohol abuse, asthma, cancer, diabetes, drug abuse, heart disease, hypertension, obesity, pain, polio, tuberculosis, and others.

6. Operational Considerations

6.1 The key to implementing universal metrics for rating scale measures will be a means of documenting, monitoring, and maintaining scale calibrations in a public forum. Many industries, including the health care industry, already have such systems in place. Instruments for measuring height, weight, temperature, blood pressure, vision, the chemistry of body fluids, etc. are designed and manufactured to measure in a particular unit within a particular range of error. There is no obvious reason why the design and manufacture of rating scale measurement systems should be less standardized than those of any other kind of measurement system, and there are many obvious reasons why they should be just as standardized. ASTM's experience as a forum in which industry representatives can negotiate the elements of measurement quality standards and standard measures makes it the natural place to situate discussions of rating scale measurement system standards.

7. Dictionary of Health Status Measures

7.1 The minimum recommended subject areas include a clinical or self-report version of each of the following:

Physical function measure
Physical function error
Physical function data quality index (model fit statistic)
Fine motor skills measure
Fine motor skills error
Fine motor skills data quality index
Cognitive function measure
Cognitive function error
Cognitive function data quality index
Psychosocial function measure
Psychosocial function error
Psychosocial function data quality index
Spiritual well-being measure
Spiritual well-being error
Spiritual well-being data quality index
Quality of life measure
Quality of life error
Quality of life data quality index
Satisfaction with services measure
Satisfaction with services error
Satisfaction with services data quality index

7.2 A comprehensive array of additional subject areas that could easily be surveyed by computer-adaptive instrument administration and analysis include:

Physical function measure
Physical function modeled error
Physical function fit-inflated error
Physical function mean square infit (information-weighted model fit)
Physical function standardized infit
Physical function mean square outfit (outlier-sensitive model fit)
Physical function standardized infit
Physical function point biserial correlation

Fine motor skill measure
Fine motor skill modeled error
Fine motor skill fit-inflated error
Fine motor skill mean square infit (information-weighted model fit)
Fine motor skill standardized infit
Fine motor skill mean square outfit (outlier-sensitive model fit)
Fine motor skill standardized infit
Fine motor skill point biserial correlation

Cognitive function measure
Cognitive function modeled error
Cognitive function fit-inflated error
Cognitive function mean square infit (information-weighted model fit)
Cognitive function standardized infit
Cognitive function mean square outfit (outlier-sensitive model fit)
Cognitive function standardized infit
Cognitive function point biserial correlation

Psychosocial function measure
Psychosocial function modeled error
Psychosocial function fit-inflated error
Psychosocial function mean square infit (information-weighted model fit)
Psychosocial function standardized infit
Psychosocial function mean square outfit (outlier-sensitive model fit)
Psychosocial function standardized infit
Psychosocial function point biserial correlation

Spiritual well-being function measure
Spiritual well-being function modeled error
Spiritual well-being function fit-inflated error
Spiritual well-being function mean square infit (information-weighted model fit)
Spiritual well-being function standardized infit
Spiritual well-being function mean square outfit (outlier-sensitive model fit)
Spiritual well-being function standardized infit
Spiritual well-being function point biserial correlation

Quality of life measure
Quality of life modeled error
Quality of life fit-inflated error
Quality of life mean square infit (information-weighted model fit)
Quality of life standardized infit
Quality of life mean square outfit (outlier-sensitive model fit)
Quality of life standardized infit
Quality of life point biserial correlation

Satisfaction with services measure
Satisfaction with services modeled error
Satisfaction with services fit-inflated error
Satisfaction with services mean square infit (information-weighted model fit)
Satisfaction with services standardized infit
Satisfaction with services mean square outfit (outlier-sensitive model fit)
Satisfaction with services standardized infit
Satisfaction with services point biserial correlation

7.3 Data quality interpretation is an art. The statistical significance of small deviations from the modeled pattern of responses grows as the number of survey questions administered increases. The meaning of numeric data quality indicators

(model fit statistics) is therefore not constant and must be interpreted in light of various contextual factors that can include patient comorbidities and the relevance of the survey questions for the patient in question, in addition to the number of questions asked. For these reasons, the array of data quality indicators listed must be made available to trained personnel for use in determining data quality and its relationship, if any, to treatment outcomes and the quality of care.

7.4 Similar arrays of subject areas will be required for each of the disease-specific measures.

ANNEXES

(Mandatory Information)

A1. DEFINITIONS AND DISCUSSION OF SCALE-FREE MEASUREMENT TERMS

A1.1 *Adaptive Measurement*—Adaptive measurement (Choppin, 1968, 1974; Wright & Douglas, 1975; Wright & Bell, 1984; Weiss, 1983; Weiss & Kingsbury, 1984; Lunz, et al., 1994) is the practical advantage that follows from Rasch measurement's capacity to take missing data into account (see *logit*). When the quantitative hypothesis (see *measurement*) is not falsified and a stable construct is identified, the amount of the variable measured by individual survey or test questions is established as constant, within the error of measurement, for the relevant population. As such, individual questions can be selected from a precalibrated item bank according to their quantitative or cultural relevance to the respondent. There is little quantitative information obtained when there is a large difference between a person's likely measure and an item's known difficulty because of the high probability that the response will be correct or mistaken (on tests of ability), or in an extreme response category (on surveys). The questions that most productively contribute to a measurement effort are those for which the probability of response is 50/50. For a person with a low measure, then, the most quantitatively relevant questions on a test or survey are those measuring at the low end of the scale; administration of questions from the middle or high ranges of the scale (more than 3 logits above the person's measure) will not contribute quantitatively useful information.

A1.1.1 Quantitative criteria alone can be used to allow a computer to automatically select questions with the goal of obtaining a measure with a particular degree of error in the fewest possible questions. Existing computer adaptive test administration software allows the administrator to require each respondent to answer a minimum number of questions, to set a maximum error level, or to require a person with a measure close to a minimum competency level to answer enough questions to clearly establish that the measure is at least an error (or more) above or below the cutoff point. Some questions can be chosen for universal administration, or the system might be set to allow the user or a rater (therapist, nurse, physician) to choose questions based on functional, diagnostic, personal, social, or cultural relevance.

A1.2 *Additivity*—Additivity (Wright, 1985, 1999), a combination of the mathematical field properties of commutativity and associativity, is the mathematical expression of what the philosopher N. R. Campbell (Campbell, 1920) called concatenation. Commutativity requires that the order in which arithmetical operations are performed on any two numbers not affect the result. So, $a + b$ must equal $b + a$. Associativity requires that the grouping of the numbers not affect the result. So, if $a = b$, then $a + a = a + b = b + b$. The additivity of rating scale and test data cannot be assumed, but must be required by specifying particular data structures, and it must be tested by evaluating the extent to which data meet the specified requirements (see *data quality*, *measurement*, *reproducibility*, *sufficiency*, and *validity*) and produce *scale-free* measures.

A1.3 *Bias Analysis*—In multifaceted measurement designs, judges may exhibit inconsistencies in their ratings relative to a particular area of performance or to a particular person observed. Bias analysis is the investigation of these inconsistencies. Linacre's Facets software (Linacre, 1994, p. 54) fits data to multifaceted models and detects bias by partitioning response residuals by element (judge-item pairs) and converting these into logit measures. The size of these logits are evaluated via statistical tests in order to find systematic patterns of inconsistency. By specifying elements for investigation, small but systematic biases can be detected when they might otherwise be lost within the unavoidable background noise present in the data. Because the amount of bias is reported as a measure with a standard error and an associated statistical significance, the practical implications of retaining or removing the bias can be evaluated. Finally, the amount of systematic bias hidden within the random noise in the data can be estimated by partitioning the unexplained error variance. This is revealed by correcting the bias logit standard deviation by its measurement error to estimate the amount of systematic error in the error variance. This is accomplished by taking the square root of the bias logit variance minus its error variance. If a bias analysis of judges results in a higher estimate of systematic bias than is indicated by the judges' root mean square error (RMSE) estimate, then a few extremely inconsistent judge-person interactions have probably caused systematic bias to be overestimated. On the other hand, if the estimate of systematic bias is lower than the RMSE, then the general consistency of the judge-person interactions is dampening the appearance of systematic biases present in the data.

A1.4 *Calibration*—Calibration is the process of testing and establishing additivity and reproducibility in a data set by evaluating its fit to a mathematical model requiring these features. Calibration is effected by employing any of several estimation and data quality evaluation methods. Calibrations are the quantitative scale values of survey questions and

response category steps; the term is also used in reference to the scale values associated with judges and other parameter estimates in multifaceted designs.

A1.5 *Concatenation (NR Campbell)*—In his landmark 1920 work, Physics, The Elements, Campbell defined measurement as the concatenation of physical unit quantities, as in the laying of a block of unit size end to end with itself, or in the piling on of identical unit weights in the pan of a balance beam. In stressing the physical nature of the units concatenated, Campbell repeated the Pythagorean fallacy of misplaced concreteness and ignored the more productive Platonic approach to the objects of measurement as ideals. As is documented by Michell (Michell, 1990), the effect of Campbell's popular definition of measurement on psychology produced lasting deleterious effects. Largely because of the mistaken notion that quantitative measurement is the sine qua non of science, psychologists first put themselves through contortions to find physical unit quantities that they could base their science on, and then they redefined measurement in a way that allowed just about any assignment of numbers to qualify as scientific. In the 1920s, L. L. Thurstone (1959) came up with a definition of measurement far closer to the Platonic ideal than Campbell's, but the cumbersomeness of his methods allowed Likert's (1932) simpler and less rigorous formulation to hold sway. Likert's methods also based measurement validity more on the familiar Pythagorean ground of survey content, meaning that the simpler method resonated better with established research mores, though the advantages of rigorous measurement were not made available (Fisher, 1992, 1994). The end effect was to remove virtually all criteria for Thurstone's "crucial experimental test": evaluating the extent to which instruments are affected in their measuring functions by the objects of measurement. With the introduction of Rasch measurement, relatively simple, flexible, and easily applied criteria for implementing Thurstone's test are available.

A1.6 *Construct*—A construct is a conceptual domain measured, such as reading ability, physical disability, or customer satisfaction. A construct is a vaguer and more general concept than the concept of the variable is. Construct validity is the most fundamental kind of validity because it asks whether the thing measured is the thing that was supposed to have been measured (Cherryholmes, 1988). Construct validity requires internally consistent data (Messick, 1975, 1981) and so depends on demonstrated statistical sufficiency and the resulting parameter separation as indicating that the question and answer interactions represented in the data are evidence that a single conversational object has dominated all of the exchanges.

A1.7 *Convergence*—This term has two closely related uses. First, the Newton-Raphson method of iteratively using each parameter to improve the estimation of the other(s) leads to convergence when the differences between subsequent estimates for each parameter decrease. Second, convergence of the estimates is made possible only to the extent that differences between the observed and modeled ratings also steadily decrease as iteration progresses (see *separability theorem*, *data quality*, *statistical consistency*, and *model fit*). Rasch measurement software frequently allows for specification of maximum numbers of iterations and convergence criteria, such as the maximum tolerable logit change from iteration to iteration, and the maximum tolerable difference between observed and expected ratings.

A1.8 *Counting*—Counting is the basic activity that measurement depends on. For the purposes of measurement, different things are fictitiously considered to be the same and are counted. No measure is ever truly objective and absolutely factual since it is always accompanied by some amount of error. Measures and counts are fictions that we heuristically use to guide us through the world. It is often said that the value of dramatic fiction in poetry, novels, the theatre, music, and art is that truths that speak of and to all of us can be said aloud in a public forum without incriminating any of us as individuals, and even though the events presented never happened exactly as portrayed. It may be that the human value of scientific fictions is not much different, qualitatively or quantitatively, from the human value of artistic fictions (Fisher, 1994, 1995). Counted correct answers or steps taken on a rating scale across several items are ordinal, raw scores, cannot be assumed to function as sufficient statistics, and must not be confused with measures.

A1.9 *Data*—Data are observations that are made in such a way as to lead to generalization. Scientific data may be either qualitative or quantitative. Ordinal data are a sufficient and necessary basis for quantification, which is inherently ratio and/or interval. Ordinal data are often comprised of counts of correct responses or of steps taken on a rating scale over multiple items (see *counting* and *raw scores*). Data never fit a model perfectly, but that fact should not lead to the use of models that make it impossible to achieve generality. Data structures are described by statistical models, but are prescribed by measurement models. Thus, statistical models are fit to data; when the model fails, it is replaced by another. In contrast, data are fit to measurement models. When the data fail to support generalization, they should be removed from the present instrument calibration effort, and explanations for the failure should be sought in the texts of the questions asked, in the response options offered, in the questioning and data entry processes, and in the possible identification of individual or groups of respondents who may have identifiable reasons for producing data of a different consistency.

A1.10 *Data Quality, Statistical Consistency, and Model Fit*—All three of these are interrelated and depend extensively on what Thurstone (1959, p. 228) called the "crucial experimental test" that establishes whether an instrument is affected in its measuring function by the object of measurement. Rasch realized a less cumbersome way of implementing Thurstone's crucial experiment when he saw that Fisher's (1922) "formalization of sufficiency nails down the ... conditions that a model must fulfill in order for it to yield an objective basis for inference" (Rasch, personal communication recorded in Wright, 1980, p. xii). Poor data quality, statistical inconsistency, and inadequate model fit all follow from a failure to pass the crucial experimental test, since the conditions of sufficiency

will not have been met.

A1.10.1 When raw scores function as sufficient statistics, the probability of success on an item consistently increases as a person's score increases, given a set test length, no matter which item is addressed, and, conversely, the probability that an item will be succeeded on increases as its score increases, given a set sample size, no matter which person is involved. When items from a test or survey fall into a stable hierarchy that persists across respondents, and when the respondents similarly take up a stable order over the items, the data are statistically consistent and likely to fit the specified model. A Rasch model is a mathematical statement of the requirement that raw scores function as sufficient statistics.

A1.10.2 With the worst quality data, convergence will not be achieved, since insufficient raw scores will not support the formulation of expectations the data can live up to. Complete lack of convergence is relatively rare. More commonly, local disturbances in the measurement process do not prevent convergence, but inflate one or more model fit statistics to a statistically significant level for some respondent(s) or item(s). These disturbances can result from data entry errors, ambiguous questions, respondent inattention, mistaken response option coding, or the administration of items irrelevant to particular respondents because of their (the items') content, extreme difficulty, or extreme easiness.

A1.10.3 Commonly employed model fit statistics include information-weighted and outlier-sensitive variance ratios. The information-weighted fit statistic, referred to as infit, is the ratio of the observed information variance to the model-expected variance; that is, each squared residual difference between expected and observed responses is weighted by the model-expected variance before being divided by it. The outlier-sensitive fit statistic, referred to as outfit, is the ratio of the sum of the squared residuals to the sample size. When data fit a model, these statistics approximate a mean square distribution, with an expected value of 1.0. For more on evaluating data quality, including methods for calculating fit statistics, see Wright and Stone (1979), Wright and Masters (1982, pp. 99-101), or Smith (2000).

A1.11 *Determinism and Probability*—Deterministic measurement models, such as that proposed by Guttman (Guttman, 1950), require all counts to be perfectly sufficient for reproducing the pattern of responses over the length of the instrument.[4] Just as a measure of eight centimeters requires that the object measured be as long as the distance from the origin to the eighth centimeter mark on a meter stick, so do deterministic models require that all abilities and attitudes perfectly conform to the items' difficulty hierarchy. The practical result of this over-rigid requirement is that much data, even most of the data in a study, are deemed unscalable (Wilson, 1989). Rasch's models for measurement are sometimes described as probabilistic Guttman models (Andrich, 1985; Brink, 1972) because of the two approaches' common focus on reproducibility and sufficiency. Although Rasch's models retain Guttman's focus on tests of reproducibility/sufficiency, their probabilistic structure makes them much more practical for use in the measurement of human abilities and attitudes than Guttman's deterministic model.

A1.12 *Dimensionality*—In so far as measurement is inherently a matter of repeating a single unit amount along a linear dimension, it is uni-dimensional. In so far as measurement is always associated with some degree of error, it is always multi-dimensional. Whether the unavoidable amounts of multi-dimensionality always present in data are enough to subvert the practical utility of the measuring device is not a question that any statistic can answer. The gross limit of unidimensionality can be seen in a signal to noise, or variation to error, ratio (see *reliability*). When the measurement effort has not resulted in any separation of those measured along a continuum of more and less, then the error is likely to be greater than the variation, and a number line we can count on for consistent indications of amount has not been drawn out. Similarly, if variation in raw scores has been produced, but the scores are so statistically inconsistent that error overwhelms that variation, then no single dimension has been delineated here, either. On the other hand, when a standard deviation of the linear logits is two or more times the error, a number line can be drawn between each statistically distinct stratum and can be relied upon to provide stable quantities. See *data quality*.

A1.13 *Equating, Cocalibration*—There are two basic methods by which instruments can be equated, or cocalibrated, so that they measure in the same quantitative unit (Wright & Stone, 1979; Masters, 1984). The method by which existing, separately developed instruments are most commonly equated is through their application to a common sample. The item calibrations from the joint administration are compared to their prior, independent calibrations on (usually larger) samples. If the common and separate calibrations correlate highly (0.85 or higher), the next step is to anchor the items at the common sample calibration's provisional values and use these anchored values to produce a separate measure from each instrument on each person in the common sample. These measures should then also produce a high correlation coefficient (0.85 or higher), and should be examined graphically to establish that the instruments produce the same measure (Altman & Bland, 1986; Bland and Altman, 1986; Masters, 1984).

A1.13.1 When an existing instrument is changed via the addition, deletion, or modification of items, enough items need to be retained from one version to the next to support equating. If new and old items are administered together in a pilot study, the resulting analysis is the same as the common sample equating. If old items are dropped in favor of new ones, the remaining items form the basis for a common item equating. In this case, the separate samples are combined as if they were one, and the new, old, and common items are analyzed together. This approach is also the basis of item bank building.

[4] Guttman's sense of reproducibility was apparently invented without influence from Fisher's notion of sufficiency, as Guttman makes no reference to Fisher. Reproducibility in Guttman's sense is more closely related to the ASTM sense of repeatability than it is to the ASTM sense of reproducibility; see ASTM standards E 456 and E 691 for further details on the contrast between repeatability and reproducibility.

A1.13.2 These equating methods will be an important part of interlaboratory instrument testing, as they establish different instruments' relative reliabilities and validities in a common framework.

A1.14 *Error*—Measurement is never perfectly precise. Because measures are always accompanied by error, they should never be stated simply as numbers, but should always be expressed as estimates bounded by a stated range of error. Similarly, because the observational data (rating scale or test scores) from which measures are estimated are never perfectly consistent, and would be suspect even if they were, measures should always be accompanied by model fit statistics (see *data quality*). The amount of error associated with a measure is most directly a function of the number of questions asked and the number of rating scale points offered as observational distinctions. Wright and Masters (Wright & Masters, 1982: 66-68) estimate measurement error in the PROX (normal approximate) procedure using the following formula:

$$SE(b_r) = \sqrt{X\,[mL/(r(mL - r))]} \quad \text{(A1.1)}$$

where $SE\ (b_r)$ is the standard error SE for person b with score r; X is a test spread expansion factor that is used to remove the effect of the dispersion of the particular items employed from the person measures and errors; and mL is the maximum score possible given the length of the test. Conversely, item calibration error is estimated in the PROX context as:

$$SE(d_i) = \sqrt{Y\,[mN/(S_i\,(mN - S_i))]} \quad \text{(A1.2)}$$

where $SE\ (d_i)$ is the standard error SE for item d with score i; Y is a sample spread expansion factor that is used to remove the effect of the dispersion of the particular sample responding from the item calibrations and errors; S_i is the score S for item i; and mN is the maximum score possible given the number of persons responding. Wright and Stone (1982, p. 89) also offer equations for estimating asymptotic standard errors in the context of an unconditional maximum likelihood procedure that does not require the assumption of a normal distribution.

A1.14.1 The fundamental relationship employed in estimating error is that of the observed and maximum possible raw scores. Given raw scores functioning as sufficient statistics, the lowest error of measurement is achieved when about half the questions asked are answered correctly (or when the probability of response in either of two adjacent rating categories is 50-50). As the observed score approaches the minimum or maximum extreme, the error goes up.

A1.14.2 Error can be inflated in some software programs (WINSTEPS, FACETS) to include the additional uncertainty introduced by unmodelled variation (statistical inconsistency).

A1.15 *Estimation Algorithms*—Where a Rasch model is a mathematical specification of an observational framework through which an intention to measure will be realized or defeated, any one of several estimation methods could conceivably be adapted to effect application of any particular model (Masters & Wright, 1984; Wright & Masters, 1982). These estimation algorithms include PROX (Cohen, 1979; Wright & Stone, 1979), a simple algebraic method of calculating approximate estimates that requires normal distributions of persons and items; PAIR (Rasch, 1980, pp. 171-172; Choppin, 1968; Wright & Masters, 1982), another simple method that estimates parameters via evaluations of pairs of items; FCON (or simply CON) (Andersen, 1973), a fully conditional algorithm that is extremely resource-intensive; and UCON (Wright & Panchapakesan, 1969; Wright & Masters, 1982), an unconditional method presenting a practical balance between simplicity and accuracy.

A1.16 *Incommensurable, Commensurable*—For the ancient Greeks, the incommensurability of the hypotenuse of a right isosceles triangle with the two other sides was a mathematical catastrophe of the first order. One legend holds that Pythagoras committed suicide upon learning of this discovery. Another version has a member of the cult killed for threatening to release the secret. The problem is that no degree of measurement precision will provide a rational number solution to the Pythagorean theorem for a right isosceles triangle, in which the square of the length of the hypotenuse is equal to the sum of the squares of the other two sides. Because the Pythagoreans and Sophists held to the notion that the objects of geometry were the figures actually drawn, the irrationality, and, hence, undrawability, of the length of this line segment threatened the basic principles of their mathematical thinking and logic.

A1.16.1 Plato solved the problem by redefining the objects of mathematics as ideals (see *concatenation*, *validity*), requiring that the results of any analysis be produced using only the compass and straightedge, and be reproducible, within some range of error, by trained person using any particular drawn figure. By redefining the objects of geometry as ideals only approximated by inherently error-prone concrete figures, the previously incommensurable line segments were transformed into commensurable measures.

A1.16.2 Copernicus, Kepler, and Galileo extended Plato's mathematization of world-measurement (geo-metry) to the heavens. Rasch's models for measurement effect basically the same Platonic shift with regard to the psychosocial world, restricting the instruments of the human sciences to those that allow the ideality of the objects of discourse to show themselves. Plato redefined the point as an indivisible line, a circle as a curve equally distant at all points from its center, and a line as an indivisible plane. Galileo based his theory of gravity on the behavior of objects on a frictionless plane. Extending this line of thinking, Rasch has similarly prompted us to redefine the variables of the human sciences mathematically. Most popular test and survey measurement methods focus on scale content, in Pythagorean fashion, resulting in incommensurable measures. Methods based on Rasch's separability theorem, in contrast, test the hypothesis that a variable is quantitative in the full mathematical sense of the term, so that a person's functional status or mathematics ability measure does not depend on ratings on or responses to a particular set of questions, but remains constant, within a range of error, no matter what particular questions were administered.

A1.17 *Instrument*—Typically connoting a particular collection of rating scale or test items, which may or may not address a common construct, this term is best used to refer to a sample of items addressing a common construct. These samples may be items collected together under a single brand name, or they

may be selected via self-, rater-, or computer-adaptive measurement strategies from a larger bank of items. Different brands of instruments shown to address a common construct are still separate instruments, but a bank of items assembled for tailored administration and measuring a single construct in a common metric may also be called an instrument, even if no two persons measured respond to a common set of questions.

A1.18 *Inter- and Intra-laboratory Testing*—Following the ASTM E 691 Standard Practice, intra-laboratory testing focuses on ruggedness testing and establishing repeatability, "the variability between independent test results obtained within a single laboratory in the shortest practical period of time by a single operator with a specific set of test apparatus using test specimens (or test units) taken at random from a single quantity of homogenous material obtained and prepared for the ILS [Inter-Laboratory Study]" (ASTM E 691, p. 4; also see ASTM E 177, E 456, E 1169, and Wernimont, 1978). Inter-laboratory testing focuses on establishing reproducibility, "the variability between single test results obtained in different laboratories, each of which has applied the test method to test specimens (or test units) taken at random from a single quantity of homogenous material obtained or prepared for the ILS" (ASTM E 691, p. 4). This document's companion Standard Practice for Conducting a Study to Determine the Precision of Test and Rating Scale Measures adapts the E 691 terminology and practice to the metrological needs of psychosocial measurement.

A1.19 *Item Response Theory and Latent Trait Theory*—Because of similarities with work in the areas of Item Response Theory (IRT) and Latent Trait Theory (LTT), Rasch's measurement theory is sometimes incorporated under these headings. No mention of these theories is found in Rasch's work. Wright, the student of Rasch's who has probably done the most to simplify, expand, and provide access to Rasch's work, details the reasons for distinguishing Rasch measurement from IRT (Wright, 1984); supplementary background supporting Wright's position is provided by Fisher (Fisher, 1994). In short, IRT is willing to forego the Rasch models' data-prescriptive parameter separation, sufficiency, and consequent capacity for supporting scale- and sample-free uniform data standards, such as those created and maintained in the form of traceable metrological reference standard metrics. IRT favors other models incorporating additional parameters that better describe some data, but achieve the superior description by sacrificing parameter separation. Thus, the area of theoretical work relevant to Rasch's models is measurement theory's principles of mathematical invariance, and metrological precision and bias.

A1.20 *Items, Item Banking*—Survey statements and test questions are commonly referred to as items in the psychometric literature. Item banking is a process of building up a pool of equated questions for adaptive administration. Hundreds or thousands of items relevant to a given construct might be equated via a series of linked common samples. The equating of existing instruments intended to measure a common variable is an exercise in item banking.

A1.21 *Levels of Measurement (Nominal, Ordinal, Interval, Ratio)*—The division of measurement into levels by S. S. Stevens (Stevens, 1946) has played a major role in stalling advances in quantitative research in the human sciences for more than 50 years (Duncan, 1984; Michell, 1990, 1997, 1999, 2000). By considering qualitative nominal and ordinal observations as levels of quantification, the word "measurement" became so vague and diluted that almost any data could be considered a solution to a measurement problem. Michell reminds us that the classical notion of measurement as the repetition of a unit quantity demands that the quantitative status of a variable be treated as an hypothesis and tested empirically. Though aspects of the classical approach to measurement remained alive in the areas of measurement theory represented in the works of Thurstone, Guttman, Loevinger, Luce & Tukey, Krantz, et al., Rasch, Wright, and others, this technically challenging literature has done little to change the research behaviors of most social scientists. Only interval and ratio measures are both quantitative and mathematical, and so only they should be called measures.

A1.22 *Logit*—The logit is the log odds unit, and has a long history of use in mathematics, statistics, and measurement. The log is the natural logarithm, or the number x for which e^x is a number greater than zero and e is the number that the function $(1+1/n)^n$ converges on as n increases. Logarithms occupy an important place in just about every area of pure or applied mathematics, as well as in the arts and sciences. The natural logarithm is often applied in statistics as a "two-stretch transformation," so called because of the way it stretches both tails of a distribution away from the center. The importance of this transformation for rating scale- or test-based measurement is that it linearizes ordinal data, making the unit amount in the tails equivalent to that in the center of the distribution. The nonlinearity of ordinal data is well documented (Merbitz, et al, 1989; Silverstein, et al., 1989; Stucki, et al., 1996; Wright & Linacre, 1989) and can be demonstrated with data from virtually any survey or test (Wright & Masters, 1982). Table A1.2 shows some logits and proportions. Notice that the logit difference between adjacent proportions increases as the extremes of the range are approached from the middle.

TABLE A1.1 Sample Data that Display the Conjoint Order Needed for Fit to a Rasch Model

	Items										
	Easy or Agreeable to Hard or Disagreeable										
Persons	1	2	3	4	5	6	7	8	9	10	Person Scores
Luc	1	0	1	0	0	0	0	0	0	0	2
Jean	0	1	1	1	0	0	0	0	0	0	3
Kevan	1	1	0	0	1	0	0	0	0	0	3
Laura	1	1	1	0	1	0	0	0	0	0	4
Alissa	1	1	1	1	0	1	0	0	0	0	5
Kevan	1	1	1	1	0	1	1	1	0	0	7
Nathan	1	1	1	1	1	1	0	0	1	0	7
Julia	1	1	1	1	1	1	1	0	1	0	8
Eleonore	1	1	1	1	1	1	1	1	0	1	9
Margaux	1	1	1	1	1	1	1	1	1	0	9
Item Scores	9	9	9	7	6	6	4	3	3	1	

TABLE A1.2 Logits from Proportions

Proportions	Logits
0.01	-4.60
0.05	-2.94
0.10	-2.20
0.15	-1.73
0.20	-1.39
0.25	-1.10
0.30	-0.85
0.35	-0.62
0.40	-0.41
0.45	-0.20
0.50	0.00
0.55	0.20
0.60	0.41
0.65	0.62
0.70	0.85
0.75	1.10
0.80	1.39
0.85	1.73
0.90	2.20
0.95	2.94
0.99	4.60

A1.22.1 Another important aspect of the logit concerns its odds component. The computation of the odds is based on the proportion P that the raw score sum is in relation to the maximum possible score given the particular items administered, and the number of partial credit or rating scale points available. Because information on the number of items and on which items is included in the calculation of the odds, the raw score metric's dependence on the particular items administered can be removed as a factor in scale calibration. Logits can be calculated by taking the natural logarithm of the ratio $P/(1-P)$ (Cohen, 1979; Wright & Stone, 1979; Linacre, 1999).

A1.23 *Mathematical Entities and Mathematical Thinking*—The ancient Greek category of *ta mathemata*, the mathematical, included anything that could be taught and learned, and which was learned through what is already known (Heidegger, 1967). The Romans translated *ta mathemata* as curriculum, losing much of the breadth and depth of the Greek term. Recovering and redeploying the Greek sense of mathematical thinking enables us to understand better how the mathematization of nature came about.

A1.24 *Measurement*—Measurement is the counting of a constant unit value repeated along a continuum of more and less. The amount expressed by the unit value is arbitrary, but must be invariant, 66 % of the time within a range of error, 95 % of the time within two errors. Whether the variable of interest is quantitative, that is, whether it exists in a form that can be meaningfully and usefully expressed as a unit amount bounded by error, is an hypothesis that must be tested by gathering data and examining them for the ordered and additive structures characteristic of quantities (Michell, 1990, 1997, 1999, 2000). Failure to falsify the quantitative hypothesis does not mean that it will not be falsified by other data at some other time; confirmation of the hypothesis is always provisional (Popper, 1965). Successful falsification of the quantitative hypothesis does not doom study of the variable to non-scientific status; science involves qualitative research of many kinds before and during quantitative research (Kuhn, 1961, 1972).

A1.25 *Metaphor in Measurement*—Counting is not the reduction of things that are not the same to things defined as the same; it is rather a metaphorical suspension of disbelief in the possibility of sameness in the name of a productive application (Ballard, 1978). Science does not, therefore, eliminate metaphor from its discourse (Black, 1962; Hesse, 1972; Kuhn, 1979; Gerhart & Russell, 1984; Rothbart, 1997; Hallyn 2000; Maasen & Weingart 2001), since metaphor permeates all linguistic concept formation (Ricoeur, 1977; Ortony, 1979; Gadamer, 1989; Lakoff & Johnson, 1980; Tracy, 1981, 1985). In fact, measurement is fundamentally metaphorical in its capacity to structure new analogies, such that the scale value (calibration) of item A is to the scale value of item B as the measure of person C is to that of person D (A:B::C:D), or A:C::B:D, etc. (Fisher, 1988, 1989, 1995, 1997, 2000).

A1.26 *Metric*—A metric is a sequence of unit amounts of arbitrary size and range. Unit size can be established according to any meaningful difference in the variable of interest. For instance, an early definition for a yard of length was set, in the thirteenth century, as the length of King Henry I's outstretched arm, from the tip of his nose to the tip of his thumb. In the metric system, temperature is defined by the difference between the freezing and boiling points of water; the former is set at zero, the latter at 100, and the difference is evenly divided, based on the rate at which a substance such as mercury expands in an enclosed volume. Similar meaningful metrics are suggested below for the variables of interest for the CPR record.

A1.27 *Missing Data*—One of the ways in which the meaning of raw score sums of ordinal data remains inextricably tied to samples of respondents and items concerns the incommensurability of these scores. Varying the items administered across respondents necessarily changes the meaning of the counts of correct answers or the sums of ratings, since the difficulty or agreeability of the questions asked will vary. When the items are calibrated to stable positions along a ruler, via tests of the hypothesis that the variable is quantitative, then measures can be derived from different collections of items by comparing the relative probabilities of the responses with the item calibrations, employing any one of several different estimation algorithms.

A1.28 *Multi-faceted Measurement*—The simplest measurement designs have only two facets and estimate only two parameters, one for the persons, abilities, performances, or objects measured, and the other for the items, or questions asked. When partial credit is awarded for partially correct responses, or when a rating or Likert scale is used on a survey, the item parameters may be estimated for each pair of adjacent steps from category to category. Additional parameters may be added to a measurement model as long as the variation within it is consistent across all the other parameters. For instance, judges may rate performance quality across several tasks according to a series of criteria (Linacre, 1989; Linacre, et al., 1994). See *Rasch analysis* for more information.

A1.29 *Ordinal Data*—See *Levels of Measurement*.

A1.30 *Population*—A population is the entire universe of a facet's elements (persons, items, judges, etc.) relevant to the measurement of a particular construct. For instance, the population of persons with disabilities who could benefit from rehabilitation treatment defines the relevant disability variable. Insofar as a hierarchy of physical disability items calibrated on any sample of persons conform with the hierarchy of items relevant to a particular population, that sample can be said to belong to the population. The population defines the variable. The determination of whether apples and oranges are being compared is an empirical matter that requires deciding if the oranges are in fact yellow apples, and the color of the apples is irrelevant, or if what matters is in fact fruit, which would make immaterial the mixing of apples and oranges.

A1.31 *Probabilistic Conjoint Measurement*—Table A1.1 shows fictional data that share a probabilistic conjoint order. This order would be deterministic and perfectly sufficient, like a gross, large-unit measure of height or weight, if every person's pattern of responses were perfectly reproduced from the total score. The order is conjoint because each facet has a stable order across all of the elements of the other. Items 1-3 are the items most likely to be succeeded on or agreed with, and item 10 is least likely to be succeeded on or agreed with, no matter whether the person in question has a high or low score. The person with the lowest score is the one least likely to succeed or agree on any item, and the persons with the highest score are most likely to succeed or agree on any item.

A1.31.1 Table A1.1 is a useful framework for demonstrating the ways in which data quality, statistical consistency, and model fit are evaluated. Imagine that another item was added, and that this item differs from the others in that the persons with the lowest scores are succeeding on, or agreeing with, it, and the persons with the highest scores are not. In its implementation of Thurstone's "crucial experimental test," a Rasch model formulates expectations based on the extent to which a conjoint order holds. These expected ratings are compared with the observed ratings, producing a residual score (the left over difference between the expected and observed). Observed ratings that are much higher or lower.

A1.32 *Quantification*—Raw scores composed of counts of rating scale steps or correct answers are not quantitative measures because they do nothing to indicate how much of the thing measured a person exhibits. Amounts are constant differences between points on a number line that any instrument measuring the thing in question will indicate. Length, for instance, is the same distance between two points, no matter whether that distance is measured in inches or centimeters. Rasch measurement makes possible the cocalibration of different instruments intended to measure the same variable, allowing each of them to express amounts in a common metric. Comparison of physical disability instruments calibrated on separate samples of widely varying sizes shows that, even under these somewhat haphazard conditions, amounts of physical disability remain remarkably constant across instruments and samples (Fisher, 1997). The point is to determine the extent to which quantities are mathematical, a pairing more often assumed than demonstrated.

A1.33 *Rasch Analysis, Measurement, and Models*—Rasch models specify the observational framework and data quality necessary for quantification. What is meant by the expression "the Rasch model" is not any particular model, of which there are many, but rather the feature of parameter separation around which Rasch models are designed (Masters & Wright, 1984). Parameter separation is so fundamental to quantification that the need to associate measurement models requiring it with a particular person's name ought soon be seen as a sad accident of history, and the descriptive appellation "Rasch" dropped.

A1.33.1 Rasch's models for measurement are typically two-faceted, with parameters prescribing the structure of data pertaining to the persons measured and the items measuring; more recently developed multi-faceted models (Linacre, 1989) are more complex, with less widely-available software.

$$nlog\left(\frac{P_{ni}}{P_{ni-1}}\right) = B_n - D_i \tag{A1.3}$$

The simplest models are for binary data, such as counts of right and wrong test answers, asserting that the logit measure (the natural logarithm of the odds that person n correctly answers item i) is equal to the difference between the ability B of person n and the difficulty D of item I. Exam proctoring is generally aimed at making sure that no external factors enter into the question and answer process to exert unwanted influences on the test results. Little or nothing, however, is usually done to check the empirical evidence that such was the case. Rasch models are tools for making such checks.

A1.33.2 In response to criticism of multiple-choice testing, the above model was expanded to specify the conditions in which a question's set of response options might themselves be scaled, so that some credit is obtained by the examinee for choosing answers that are partly, but not entirely, correct, but not entirely wrong, either. These models are known as partial credit models (Masters, 1982; Wright & Masters, 1984).

$$nlog\left(\frac{P_{nij}}{P_{nij-1}}\right) = B_n - D_{ij} \tag{A1.4}$$

Here, the extra difficulty associated with taking j steps across item i's response options is estimated as part of the item's difficulty D. Because multiple choice tests rarely employ the same set of response options across items, each item has its own particular step parameter estimates. Another application for this model became apparent when it was realized that some surveys and assessments might employ items addressing a common construct, but which have different numbers of rating scale points, or different definitions for a common number of rating options. In these cases, items can be grouped together so that a common rating scale model, is applied to items with the same rating structure.

$$nlog\left(\frac{P_{nik}}{P_{nik-1}}\right) = B_n - D_i - J_k \tag{A1.5}$$

Here the difficulty of taking step k on rating scale J is not associated with an individual item, but is estimated across items sharing the same rating options.

A1.33.3 This extension of the binary model via the subtraction of a third parameter suggests that further facets might be modeled into the measurement process (Linacre, 1989), such as the extra difficulty posed by harsh judges or other uncontrollable environmental factors influencing the measurement outcome. Accordingly, a simple many-faceted Rasch model takes the form of:

$$nlog\left(\frac{P_{nimk}}{P_{nimk-1}}\right) = B_n - D_i - C_m - J_k \quad \text{(A1.6)}$$

where the only addition to the rating scale model is a parameter for the difficulty *C* presented by judge *m*. Additional parameters can be added as needed, as long as consistent linear relationships hold across all of the facets modeled. For instance, items might be modeled to apply to several different tasks. Furthermore, partial credit-like variations to these models can be added to allow, for instance, each judge to have individual interpretations of the rating scale. Additional models test for parameter separation in Poisson count, ranking, paired comparison, and other kinds of data.

A1.34 *Raw Score*—A raw score is a sum of ratings made in response to survey or assessment questions, or a count of correct responses to test questions. By far the commonest approaches to educational and psychological measurement treat raw scores as measures with no evaluation of their mathematical capacity to support the inferences made. Raw scores are commonly assumed to be sufficient statistics, and in fact must be, to be meaningful. Tests of fit to a Rasch model are no more than tests of sufficiency (Andersen, 1977).

A1.35 *Reliability*—Table A1.3 shows the relations that hold among traditional reliability coefficients and the elements of alternative statistics based on the ratio of variation to error, or signal to noise.

A1.35.1 The ceiling effect found in the traditional reliability coefficient, alpha, is overcome in the separation statistic, *G*, which is the ratio of the standard deviation to the error. Another useful indicator of reliability is the number of strata - ranges along the measurement continuum with centers separated by at least three errors - delineated. Scales with more strata plainly present more opportunities for making statistically significant distinctions than those with fewer strata. Because error terms are heavily dependent on sample size, instruments with many items can produce measures with unbelievably low errors, and items will calibrate with similarly too-low-to-be-true errors when administered to very large samples. These situations can result in reliabilities approaching 1.0 and errors less than one tenth of the logit equivalence of a one-unit change in the raw score. When this happens, more conservative and realistic estimates of the error and reliability terms can be made from the logit equivalence of a one-half unit change in the raw score. Since the sizes of the error and the raw score-logit relation are larger at the extremes of the measurement continuing than in the middle, this variation must be included when revising error and reliability estimates.

TABLE A1.3 Reliability Statistics

NOTE—See entries concerning *Reliability* and *Root Mean Square Error* for more information.

SD/Error (SA/SE)	= G	Reliability $G^2/(1 + G^2)$	Strata[A] (4G + 1)/3	% Variation Not Due to Error/ Due to Error
1/1	1	0.500	1.67	50/50
1.53/1	1.53	0.700[B]	2.37	60/40
2/1	2	0.800	3.0	67/33
3/1	3	0.900	4.3	75/25
4/1	4	0.941	5.67	80/20
5/1	5	0.962	7.0	83/16
6/1	6	0.973	8.3	86/14
7/1	7	0.980	9.67	88/12
8/1	8	0.985	11.0	89/11
9/1	9	0.988	12.3	90/10
10/1	10	0.990	14.0	91/9

[A] Ranges in the measurement continuum with centers separated by at least three error terms; see Wright & Masters, 1982, p. 92.
[B] Commonly recommended minimum.

A1.35.2 Alpha is the statistic typically used to represent reliability. As alpha increases, constant units of variation in the ratio of variation to error are represented by smaller and smaller changes in alpha. Because psychosocial measurement practice rarely focuses on deliberately asking questions that vary in the amounts of the amounts of the variable they measure, much research has settled for measurement reliabilities less than .80. General unawareness of the relation of variation to error hidden within alpha has perhaps contributed to complacency in this regard. This situation should change as the variation/error relation becomes better known, as construct theories become increasingly employed in instrument design, and as the additional statistical power associated with higher reliability is better appreciated.

A1.35.3 The modeled relationships among measurement variation, error, and reliability, on the one hand, and the number of rating scale points and items employed on an instrument, on the other, is displayed in the following nomograph (see Fig. A1.1, reproduced from Linacre, 1993). This diagram is a tool that can be used to guide instrument and experimental designs. The angled line descending from the desired reliability is traced down until it intersects the horizontal line associated with the expected measurement standard deviation. The vertical line closest to this point is then traced straight down to find the expected error of measurement and the number of items needed to obtain this reliability for designs utilizing different numbers of response options.

A1.36 *Repeatability*—See *Inter- and Intra-Laboratory Studies*. "The variability between independent test results obtained within a single laboratory in the shortest practical period of time by a single operator with a specific set of test apparatus using test specimens (or test units) taken at random from a single quantity of homogenous material obtained and prepared for the ILS [Inter-Laboratory Study]" (ASTM E 691, p. 4; also see ASTM E 177, E 456, E 1169, and Wernimont, 1978).

A1.37 *Reproducibility*—See *Inter- and Intra-Laboratory Studies*. "The variability between single test results obtained in different laboratories, each of which has applied the test method to test specimens (or test units) taken at random from a single quantity of homogenous material obtained or prepared for the ILS" (ASTM E 691, p. 4). Not to be confused with

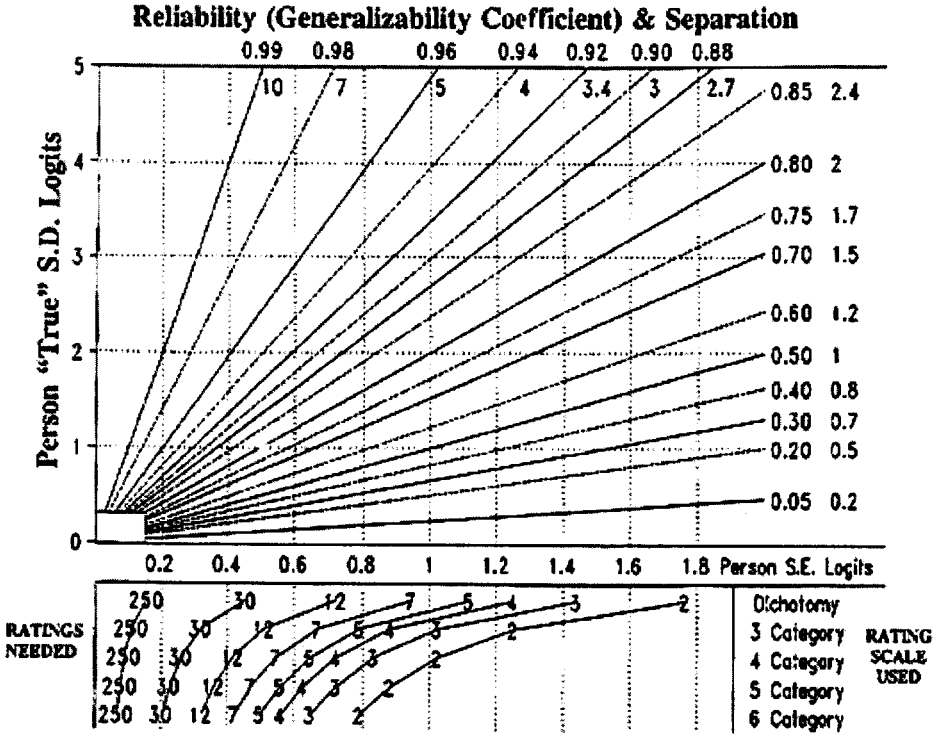

FIG. A1.1 Rasch Generalizability Theory Nomograph (Linacre, 1993; see *Reliability*)

Guttman's (1950) sense of reproducibility, which has much in common with Fisher's sufficiency and Feinstein's transparency.

A1.38 *Root Mean Square Error (RMSE)*—In order to interpret what a quantitative variable is, and to determine the extent to which a ruler that takes up a single direction has been calibrated, we need to know whether the questions asked on the test or survey are spread out along a continuum of more and less. An efficient method for evaluating the separation of the items and of the persons along the variable proceeds as follows (Wright & Masters, 1982). First, determine the amount of variation in the estimates that is due to error (the mean square error, *MSE*):

$$MSE_I = \sum_{i=1}^{L} s_i^2 / L \quad \text{(A1.7)}$$

by summing the squared item errors s (calculated as SE in Error) and dividing by the number of items administered L (for a fixed-length test or survey). Second, remove the *MSE* from the original variance of the item estimates:

$$SA_I^2 = SD_I^2 - MSE_I \quad \text{(A1.8)}$$

producing an adjusted variance SA indicating how much variation is not due to error. Third, the *RMSE* is calculated by taking the square root of the *MSE*:

$$SE_I = \sqrt{MSE_I} \quad \text{(A1.9)}$$

and this is used in the further estimation of separation and reliability.

A1.39 *Sample, Sample Size*—It is rare for any set of measures to comprise an entire population, and it concomitantly is rare for any given set of questions to comprise the entire universe of possible questions. Thus questions and responses are always samples from potentially infinite universes. One way of stating the purpose of measurement is that it aims to provide structural support for generalizations about populations and future samples from one single historical sample. In other words, measurement helps us learn how to predict and manage future measures on the basis of the repeatable and reproducible consistencies exhibited in past measures. Examined in fine enough detail, any set of measures on any variable (length, weight, temperature, volts, etc.) from any instrument are ephemeral historical events that will never happen again in exactly the same way, that is, they are never perfectly repeatable or reproducible. Rasch (1980: 115) points out that admitting this is the same thing as saying that it is the parameters in a probability distribution, and not the observed values, that take the form of natural laws. Thus, his epistemological concept of "specific objectivity" extends the mathematization of nature to the mathematization of human being.

A1.40 *Scale-free, Scale-dependent*—Counts of correct answers on a test, or sums of ratings from a survey, are scale-dependent in that their meaning changes if the number of questions asked, or the number of responses obtained, varies across examinees or respondents. Scale-dependence is a characteristic of ordinal data and can be graphically portrayed in the form of nonlinear plots of person scores from two halves of the same test, or of item scores from two halves of the respondent sample. Scale-free versions of these same plots will be linear. The terms "sample-free" and "scale-free" derive from

Thurstone's (1926, 1928) requirement of what is basically metrological repeatability: that instrument functioning not be affected by the object(s) of measurement, and that an object's measure not be affected by the particular instrument employed. Measures are never fully scale-free, and instrument calibrations are never fully sample-free, since all estimates imply some degree of error (they are not perfectly precise or accurate) and the ordinal, pre-mathematical observations from which the estimates are derived are never perfectly consistent. It is therefore important to incorporate error and data consistency statistics into any measurement application.

A1.41 *Separability Theorem, Parameter Separation*—Rasch stated the separability theorem (1980: 178; also see Rasch, 1980: 122, and Rasch, 1977) as follows: "While estimating the item parameters $\epsilon_1,\ldots,\epsilon_k$ we may eliminate the parameters $\xi_1,\ldots,\xi_n$, having, as it were, replaced them by the observed marginals $a_1.,\ldots,a_n$. And the other way around: While estimating $\xi_1,\ldots,\xi_n$ we may eliminate $\epsilon_1,\ldots,\epsilon_k$. Finally, while controlling the model we may eliminate both sets of parameters." Parameter separation does not occur without parameter convergence. That is to say, the instrument's questions and the person's responses must both participate equally in the same variable, the ability, attitude, performance, or thing measured, for the measures to be independent of the instrument selected, and for the instrument's item calibrations to be independent of the sample measured. As shown by Wright (1999), Rasch's separability theorem is the most efficient and applicable formulation to date of the requirements for measurement. Wright (1999) connects the separability theorem to a wide selection of theoretical approaches to measurement. These include Campbell's (1920) sense of concatenation, Fisher's (1922) sense of sufficiency, Thurstone's (1928) "crucial experimental test" for invariant instrument functioning, Guttman's (1944) sense of reproducibility (which is more similar to metrological repeatability than to metrological reproducibility), and to Luce and Tukey's (1964) conjoint additivity.

A1.42 *Software*—There are a variety of different software packages capable of constructing fundamental measures and implementing Rasch models. They include ASCORE (Andrich, et al., 1990); BIGSTEPS (Wright & Linacre, 1986-96); CONQUEST (Wu, et al., 1998), FACETS (Linacre, 1989-96); IPARM (Smith, 1991); LOGIMO (Kelderman & Steen, 1988); MATS (Wilson, 1996); MFORMS (Schulz, 1988); MULTILOG (Thissen, 1991); OPLM (Verhelst, 1993); PML (Gustafsson, 1979); QUEST (Adams & Kboo, 1995); RASCAL (ASC, 1996); RSP (Glas & Ellis, 1995); RUMM (Andrich, et al. 2000); SCALE (Ludlow, 1992); WINSTEPS (Wright & Linacre, 1997-2000); and others. The capacity of these packages to implement different estimation methods and to specify different Rasch models varies. The best of them offer a wide variety of analysis control options and extensive tabular and graphic output. There are few, if any, studies comparing the results of these packages' analyses on common data.

A1.43 *Specific Objectivity*—Specific objectivity (Rasch, 1977; van der Linden, 1994) is achieved when data satisfy Rasch's separability theorem. The qualifier specific refers to the fact that data satisfying the separability theorem are objective in the sense that they have converged on and separated from a common object within a frame of reference defined by the populations of persons and questions involved. The convergence and separation are never perfect in that they are always accompanied by error and statistical inconsistencies. When, however, there is more variation in the scale values than can be accounted for by error and inconsistency (see *reliability*), the parameters separate, and specific objectivity is obtained.

A1.44 *Standard, Standardized*—These words are almost always used in reference to a common instrument content, in the contexts of standardized health surveys and tests of knowledge or ability. Because the meanings of summed ratings or counts of right answers depend on a common number of questions per respondent (and vice-versa), standardized instruments are usually collections of questions deemed relevant to all members of a particular population. Any time the population changes, or the instrument is deemed too long or short, or too detailed or vague, a new instrument, with its own distinct, idiosyncratic, arbitrary and incommensurable metric is devised. Hence, we have the recent proliferation of surveys and assessments in health care.

A1.44.1 There are, however, alternative methods of standardization that are oriented toward the construct, not the content. Construct-oriented standardization produces non-arbitrary metrics that do not depend on the particular questions asked for their unit size. For the purposes of the CPR, standardized health measurement instruments would most advantageously be defined as those calibrated to measure in universal, non-arbitrary scale-free metrics. Besides replacing ordinal, scale-dependent metrics with interval, scale-free metrics expressed in a standard, common quantitative language, the construct-oriented sense of standards in measurement includes quality standards. Error and data consistency terms should be estimated for every item calibration and person measurement. Statistical summaries (means and standard deviations) of these quality indicators should then be used as estimates of overall reliability and validity.

A1.45 *Sufficiency*—Rasch studied with Ronald Fisher in London in 1935-6 and came away impressed with the singular importance of the concept of sufficiency (Wright, 1980: xi-xii; Andrich, 1979). Sufficient statistics are those that extract all available information from the data, such that the statistic is equivalent in subsequent estimations to the original data (Fisher, 1922). Arnold (Arnold, 1985) and Hall, Wijsman, and Ghosh (1965) show that "that the set of invariant rules based on a sufficient statistic is an essentially complete subclass of the class of invariant rules." Wright (1980: 193) points out that estimation is actually the second half of a story that ought to begin with substantiation of the assumption that the estimators are sufficient statistics. Or, as Michell (Michell, 1990, 1997, 1999, 2000) puts it, whether or not a variable is quantitative is an hypothesis that needs to be tested via experimental procedures that put the hypothesis at risk for falsification. Andersen (1977) details the role of sufficiency in Rasch's probabilistic conjoint models of fundamental measurement.

A1.46 *Targeting*—Targeting is a lack of floor and/or ceiling effects in measurement. The quantitative relevance of a test or survey is shown by the extent to which the distributions of item and person estimates overlap, with few or no perfect extreme scores. Wright and Stone (1979: 8) show several possible ways in which targeting affects measures. Two people of different abilities or attitudes may, depending on the targeting of the questions asked, (*1*) be as different as possible, obtaining minimum and maximum extreme measures; (*2*) be as alike as possible, obtaining the same minimum, intermediate, or maximum measure; or (*3*) be separated by different responses on one or more items that enter into the space between them along the variable. Better targeted tests measure with lower error, and hence have higher reliability. They are also more interpretable, since there is no range in the measurement continuum not associated with specific item content for which respondents in that range have a 50/50 chance of correctly (or agreeably, etc.) responding.

A1.47 *Transparency*—Feinstein (Feinstein, 1984) offers what appears to be another (in addition to Guttman's) independent invention of Fisher's notion of sufficiency. The choice of the word "transparent" to describe this feature of measurement follows from the capacity to "look through" the raw score to the composite ratings producing it. Feinstein despairs of making many rating scale instruments transparent as he lacks a sense of the way conjoint order and probability (see *probabilistic conjoint measurement*) can be used to found an objective basis for inference, as Rasch has done.

A1.48 *Unit of Measurement, Least Observable Difference, Least Meaningful Difference*—Research on metrological reference standards for the human sciences will focus on refining units of measurement based on observable differences in performance, ability, judgement, etc. The smallest differences that can be observed consistently across performances, criteria, and/or judges will be important theoretical developments that may often lead to improved instrumentation. These least observable differences will not, however, always be meaningful, in the sense of spanning a quantitative range that exceeds the range of error.

A1.49 *Validity, Construct and Content*—Plato restricted the instruments of geometry to the compass and straightedge in order to allow the ideality of geometrical figures to dominate their study. In contrast, the Pythagoreans and Sophists of Plato's day held that the actual figures drawn were the object of study. Plato's definitions of the point as an indivisible line, of the line as an indivisible plane, etc. won out in the history of science because of the parsimonious, elegant, and simple solutions this mathematical approach offered. Where Plato was more interested in constructs, the Pythagoreans and Sophists focused on content.

A1.49.1 Throughout the history of science, however, Plato's idealistic sense of objectivity has been routinely misinterpreted or ignored by philosophers and social scientists in favor of the Pythagorean/sophistic concrete sense of objectivity (Fisher, 1992, 1994). Michell (1990) documents the history of this fallacy of misplaced concreteness in 20th century psychological measurements and shows how this led to our current sense that the objectivity of rating scale- and test-based measurement depends more on content validity than it does on construct validity. As a result, most psychosocial measurement today depends on the manipulation of content-dependent, nonlinear scores.

A1.49.2 The qualitative methods (phenomenology, hermeneutics, deconstruction, post-structuralism, etc.) that are increasingly popular in the human sciences are creating the intellectual context for a shift away from the emphasis on content validity, toward a greater focus on construct validity (Cherryholmes, 1988; Fisher, 1994; Woodcock, 1999). Because reliability and statistical consistency could conceivably occur by accident from randomly chosen questions and respondents, the conjoint order of the persons and items must be required to make sound theoretical sense before it can be considered valid (Stenner, 1982).

A1.50 *Variable*—General usage allows any kind of categorical distinction to be considered a variable. With regard to rating scale- or test-based measures, a variable is something able to vary along a continuum of more and less that can be counted on to function as a number line.

A2. BIBLIOGRAPHY ON MEASUREMENT AND PHILOSOPHY

A2.1 *Bibliography:*

Ackermann JR. Data, instruments, and theory: a dialectical approach to understanding science. Princeton, New Jersey: Princeton University Press, 1985.

Altman DG, Bland JM. Measurement in medicine: The analysis of method comparison studies. Statistician 1983;32:307-17.

Andersen EB. Sufficient statistics and latent trait models. Psychometrika 1977;42(1):69-81.

Andersen EB. What Georg Rasch would have thought about this book. In: Fischer GH, Molenaar IW, editors. Rasch models: foundations, recent developments, and applications. New York, New York: Springer-Verlag, 1995: 383-90.

Andrich D. A rating formulation for ordered response categories. Psychometrika 1978;43:357-74.

Andrich D. Interview with Georg Rasch, Laesoe, Denmark, June, 1979.

Andrich D. An elaboration of Guttman scaling with Rasch models for measurement. In: Tuma NB, editor. Sociological methodology 1985. San Francisco: Jossey-Bass, 1985: 33-80.

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

Arnold SF. Sufficiency and invariance. Statistics & Probability Letters 1985; 3 September:275-9.

ASTM Committee E-11 on Statistical Methods. Standard practice for conducting an interlaboratory study to determine the precision of a test method. E 691-92. Annual Book of ASTM Standards. Philadelphia, PA: ASTM, 1992.

Bailar BA. Quality issues in measurement. International Statistical Review 1985;53(2):123-39.

Barrett W. The illusion of technique: a search for meaning in a technological civilization. Garden City, New York: Anchor Books, 1979.

Black M. Models and metaphors. Ithaca, NY: Cornell University Press, 1962.

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.

Brink NE. Rasch's logistic model vs the Guttman model. Educational and Psychological Measurement 1972;32:921-7.

Brogden HE. The Rasch model, the law of comparative judgment and additive conjoint measurement. Psychometrika 1977;42:631-4.

Brown HI. Perception, theory and commitment: the new philosophy of science. Chicago, Illinois: University of Chicago Press, 1977.

Bruninks R, Woodcock R, Hill B, Weatherman R. Development and standardization of the scales of independent behavior. Allen, TX: DLM Teaching Resources, 1985.

Bud R, Cozzens SE, editors. SPIE Institutes, Vol. 9: Invisible Connections: Instruments, Institutions, and Science, Potter RF, editor. Bellingham, WA: SPIE Optical Engineering Press, 1992.

Cella DF, Lloyd SR, Wright BD. Cross-cultural instrument equating: Current research and future directions. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials (2d edition). New York, New York: Lippincott-Raven, 1996: 707-15.

Chang W, Chan C. Rasch analysis for outcomes measures: Some methodological considerations. Archives of Physical Medicine and Rehabilitation 1995;76(10):934-9.

Cherryholmes C. Construct validity and the discourses of research. American Journal of Education 1988;96(3):421-57.

Choppin B. An item bank using sample-free calibration. Nature 1968;219:870-2.

Cohen L. Approximate expressions for parameter estimates in the Rasch model. British Journal of Mathematical and Statistical Psychology 1979;32:113-20.

Crowther CS, Batchelder W H, Hu X. A measurement-theoretic analysis of the fuzzy logic model of perception. Psychological Review 1995 April;102(2):396-408.

Douglas G, Wright BD. The two-category model for objective measurement. MESA Psychometric Laboratory Research Memoranda, vol. 34. Chicago: Dept of Education, University of Chicago, 1986.

Duncan OD. Notes on social measurement: historical and critical. New York: Russell Sage Foundation, 1984.

Englehard G. The measurement of writing ability with a many-facet Rasch model. Applied Measurement in Education 1992;5(3):171-91.

Englehard G, Osberg D. Constructing a test network with the Rasch measurement model. Applied Psychological Measurement 1983;7(3):283-94.

Fisher AG, Bryze KA, Granger CV, Haley SM, Hamilton BB, Heinemann AW, et al. Applications of conjoint measurement to the development of functional assessments. International Journal of Educational Research 1994;21(6):579-93.

Fisher RA. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London, A 1922; 222:309-68.

Fisher WP Jr. Recent developments in the philosophy of science pertaining to problems of objectivity in measurement. Rasch Measurement Transactions 1988;2(2):1-3.

Fisher WP Jr. Truth, method, and measurement: The hermeneutic of instrumentation and the Rasch model [dissertation]. Dissertation Abstracts International 1988;49:0778A. Chicago, Illinois: University of Chicago. 376 pages, 23 figures, 31 tables.

Fisher WP Jr. Bettelheim's test. Rasch Measurement Transactions 1991;5(3):164-5.

Fisher WP Jr. Scientific measurement: The supremacy of ideals. Rasch Measurement Transactions 1991;5(2):139-40.

Fisher WP Jr. Objectivity in measurement: A philosophical history of Rasch's separability theorem. In: Wilson M, editor. Objective measurement: theory into practice. Vol. I. Norwood, New Jersey: Ablex Publishing Corporation, 1992: 29-58.

Fisher WP Jr. Stochastic resonance and Rasch measurement. Rasch Measurement Transactions 1992;5(4):186-7.

Fisher WP Jr. Technology, authenticity, and educational measurement [paper]. American Educational Research Association. San Francisco, California, April, 1992.

Fisher WP Jr. Measurement-related problems in functional assessment. The American Journal of Occupational Therapy 1993;47(4):331-8.

Fisher WP Jr. Counting as metaphor. Rasch Measurement Transactions 1994;8(2):358.

Fisher WP Jr. Quality, quantity, and invariance. Rasch Measurement Transactions 1994;8(1):341.

Fisher WP Jr. The Rasch debate: Validity and revolution in educational measurement. In: Wilson M, editor. Objective measurement: theory into practice. Vol. II. Norwood, New Jersey: Ablex Publishing Corporation, 1994: 36-72.

Fisher WP Jr. Fuzzy truth and the Rasch model. Rasch Measurement Transactions 1995;9(3):442.

Fisher WP Jr. Truth from fiction. Rasch Measurement Transactions 1995;8(4):401.

Fisher WP Jr. The Rasch alternative. Rasch Measurement Transactions 1996 Winter;9(4):466-7.

Fisher WP, Jr. Physical disability construct convergence across instruments: towards a universal metric. Journal of Outcome Measurement 1997; 1(2):87-113.

Fisher WP, Jr. A research program for accountable and patient-centered health status measures. Journal of Outcome Measurement 1998; 2(3):222-39.

Fisher WP, Jr. Foundations for health status metrology: The stability of MOS SF-36 PF-10 calibrations across samples. Journal of the Louisiana State Medical Society 1999; 151(11):566-78.

Fisher WP, Jr. Mathematics, measurement, metaphor, metaphysics. Parts I and II. Journal of Theoretical & Philosophical Psychology 2000; in review.

Fisher WP, Jr. Objectivity in psychosocial measurement: what, why, how. Journal of Outcome Measurement 2000; 4(2):527-63.

Fisher WP, Jr. Survey design recommendations. Popular Measurement 2000; 3(1):58-9.

Fisher WP, Jr. Measuring outcomes in rehabilitation. In: Issues in Low Vision Rehabilitation Service Delivery, Policy and Funding, edited by Massof RW, Lidoff L, 159-83. New York, NY: AFB Press, 2001.

Fisher WP, Jr. Consequences of mathematical metaphysics for metrology in the human sciences. In: Renascent pragmatism: studies in law and social science, edited by Morales A, in preparation. Brookfield, VT: Ashgate Publishing Co., 2001.

Fisher WP, Jr, Eubanks RL, Marier RL. Equating the MOS SF36 and the LSU HSI physical functioning scales. Journal of Outcome Measurement 1997; 1(4):329-62.

Fisher WP Jr, Fisher AG. Applications of Rasch analysis to studies in occupational therapy. Physical Medicine and Rehabilitation Clinics of North America 1993;4(3):551-69.

Fisher WP Jr, Harvey RF, Kilgore KM. New developments in functional assessment: Probabilistic models for gold standards. NeuroRehabilitation 1995;5(1):3-25.

Fisher WP Jr, Harvey RF, Taylor P, Kilgore KM, Kelly CK. Rehabits: A common language of functional assessment. Archives of Physical Medicine and Rehabilitation 1995;76:113-22.

Fisher WP Jr, Wright BD, editors. Applications of Probabilistic Conjoint Measurement. International Journal of Educational Research 1994;21(6):557-664.

Fisher WP Jr, Wright BD. Introduction to probabilistic conjoint measurement theory and applications. International Journal of Educational Research 1994;21(6):559-68.

Gadamer H-G, 1960. Truth and method. Translated by Weinsheimer J, Marshall DG. Second revised edition. New York: Crossroad, 1989.

Gerhart M, Russell A. Metaphoric process: the creation of scientific and religious understanding. Fort Worth, Texas: Texas Christian University Press, 1984.

Gonin R, Lloyd SR, Cella DF. Establishing equivalence between scaled measures of quality of life. Quality of Life Research 1996:20-6.

Green J. Problems in the use of outcome statistics to compare health care providers. Brooklyn Law Review 1992 Spring;58:55-73.

Guttman L. The basis for scalogram analysis. In: Stouffer SA, et al., editors. Studies in social psychology in World War II. volume 4: measurement and prediction. New York: Wiley, 1950: 60-90.

Hacking I. Representing and intervening: introductory topics in the philosophy of natural science. Cambridge: Cambridge University Press, 1983.

Hall WJ, Wijsman RA, Ghosh JK. The relationship between sufficiency and invariance with applications in sequential analysis. Annals of Mathematical Statistics 1965; 36:575-614.

Harvey RF, Lambert RW. Rating scale analysis of a functional assessment instrument in physical rehabilitation [abstract]. Archives of Physical Medicine and Rehabilitation 1987;68:583-4.

Health care reform: 'report cards' are useful but significant issues need to be addressed. GAO/HEHS-94-219. Washington, DC: General Accounting Office, 1994.

Hedges LV. How hard is hard science, how soft is soft science? The empirical cumulativeness of research. Journal of the American Psychological Association 1987 May;42(5):443-55.

Heidegger M. What is a thing? Barton WB Jr, Deutsch V, trans. South Bend, Indiana: Regnery/Gateway, 1967.

Hess RK. Measuring school test scores [unpublished ms.]. International Objective Measurement Workshop. Berkeley, California, April, 1995.

Hesse M. Models and analogies in science. Notre Dame, Indiana: Notre Dame University Press, 1970.

Hesse M. In defence of objectivity. Proceedings of the British Academy 1972;58:275-92.

Holton G. Thematic origins of scientific thought. revised ed. Cambridge, Massachusetts: Harvard University Press., 1988.

Hunter JS. The national system of scientific measurement. Science 1980 November;210(21):869-74.

Hunter JS. Measurement error. In: Encyclopedia of Statistics. New York: John Wiley & Sons, 1986: 378-81.

Ikegami N. Functional assessment and its place in health care. New England Journal of Medicine 1995;332(9):598-9.

Kempen GI J M, Myers AM, Powell LE. Hierarchical structure in ADL and IADL: Analytical assumptions and applications for clinicians and researchers. Journal of Clinical Epidemiology 1995;48(11):1299-305.

Kline M. Mathematics: the loss of certainty. Oxford: Oxford University Press, 1980.

Kockelmans JJ, Kisiel TJ, editors. Phenomenology and the natural sciences: essays and translations, Wild J, editor. Northwestern University Studies in Phenomenology and Existential Philosophy. Evanston, Illinois: Northwestern University Press, 1970.

Krantz DH, Luce RD, Suppes P, Tversky A. Foundations of measurement. Volume 1: Additive and polynomial representations. New York: Academic Press, 1971.

Kuhn TS. The function of measurement in modern physical science. Isis 1961;52(168):161-93.

Kuhn TS. The structure of scientific revolutions. Chicago: University of Chicago Press, 1970.

Lakatos I, Musgrave A, editors. Criticism and the growth of knowledge. Proceedings of the International Colloquium in the Philosophy of Science, London, 1965, vol. 4. Cambridge: Cambridge University Press, 1970.

Lakoff G, Johnson M. Metaphors we live by. Chicago: University of Chicago Press, 1980.

Latour B. Science in action: how to follow scientists and engineers through society. New York, New York: Cambridge University Press, 1987.

Latour B, Woolgar S. Laboratory life: the social construction of scientific facts. Beverly Hills, California: Sage, 1979.

Laudan L. Progress and its problems: towards a theory of scientific growth. Berkeley, California: University of California Press, 1977.

Likert R. A technique for the measurement of attitudes. Archives of Psychology 1932; 140:5-55.

Linacre JM. Many-facet Rasch measurement. Chicago: MESA Press, 1989.

Linacre JM. Rasch generalizability theory. Rasch Measurement Transactions 1993;7(1):283-4.

Linacre JM. Facets Rasch analysis computer program. Chicago: MESA Press, 1994.

Linacre JM. Understanding Rasch measurement: estimation methods for Rasch measures. Journal of Outcome Measurement 1999; 3(4):382-405.

Linacre JM, Englehard G, Tatum DS, Myford CM. Measurement with judges: Many-faceted conjoint measurement. International Journal of Educational Research 1994;21(6):569-77.

Loevinger J. Person and population as psychometric concepts. Psychological Review 1965;72(2):143-55.

Lord FM. Small N justifies Rasch model. In: Weiss DJ, editor. New horizons in testing: latent trait test theory and computerized adaptive testing. New York, NY: Academic Press, Inc., 1983: 51-61.

Luce RD, Tukey JW. Simultaneous conjoint measurement: A new kind of fundamental measurement. Journal of Mathematical Psychology 1964;1(1):1-27.

Ludlow LH. A strategy for the graphical representation of Rasch model residuals. Educational and Psychological Measurement 1985;45:851-9.

Ludlow LH. Graphical analysis of item response theory residuals. Applied Psychological Measurement 1986;10:217-29.

Lunz ME, Bergstrom BA, Gershon RC. Computer adaptive testing. International Journal of Educational Research 1994;21(6):623-34.

Lunz ME, Wright BD, Linacre JM. Measuring the impact of judge severity on examination scores. Applied Measurement in Education 1990;3/4:331-45.

Lynch M. Discipline and the material form of images: An analysis of scientific visibility. Social Studies of Science 1985;15(1):37-66.

Mandel J. The analysis of interlaboratory test data. ASTM Standardization News 1977 March;5:17-20, 56.

Mandel J. Interlaboratory testing. ASTM Standardization News 1978 December;6:11-2.

Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982;47:149-74.

Masters GN. Common-person equating with the Rasch model. Applied Psychological Measurement 1985 March;9(1):73-82.

Masters G, Wright BD. The essential process in a family of measurement models. Psychometrika 1984;49:529-44.

Mayes P, Perez A, Mipro RC Jr, , Fisher WP Jr. Comparison of Functional Independence Measure data from the Uniform Data System and the Louisiana Rehabilitation Institute [abstract]. LSU Deparment of Medicine, Section of Physical Medicine & Rehabilitation, Residents' Research Day. New Orleans, LA: LSU School of Medicine, June, 1996.

McArthur DL, Casey K, Morrow TJ, et al. Partial-credit modeling and response surface modeling of biobehavioral data. In: Wilson M, editor. Objective measurement: theory into practice. Norwood, New Jersey: Ablex, 1992: 109-20.

McArthur DL, Cohen M, Schandler S. Rasch analysis of functional assessment scales: An example using pain behaviors. Archives of Physical Medicine and Rehabilitation 1991;72:296-304.

McHorney CA. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. [Review] [102 refs]. Annals of Internal Medicine 1997; 127(8 Pt 2) Oct 15:743-50.

Mendelsohn E. The social locus of scientific instruments. In: Bud R, Cozzens SE, editors. Invisible connections: instruments, institutions, and science. Bellingham, WA: SPIE Optical Engineering Press, 1992: 5-22.

Merbitz C, Morris J, Grip J. Ordinal scales and the foundations of misinference. Archives of Physical Medicine and Rehabilitation 1989;70:308-12.

Messick S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist 1975 October;30:955-66.

Messick S. Constructs and their vicissitudes in educational and psychological measurement. Psychological Bulletin 1981;89:575-88.

Michell J. Measurement scales and statistics: A clash of paradigms. Psychological Bulletin 1986;100:398-407.

Michell J. An introduction to the logic of psychological measurement. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1990.

Michell J. Measurement in psychology: A critical history of a methodological concept. Cambridge: Cambridge University Press, 1999.

Michell J. Normal science, pathological science and psychometrics. Theory & Psychology 2000; 10(5) October:639-67.

Myford CM. The nature of expertise in aesthetic judgment: Beyond inter-judge agreement [dissertation]. Dissertation Abstracts International 1989;50:3562A. Chicago, Illinois: University of Chicago.

Ormiston G, Sassower R. Narrative experiments: the discursive authority of science and technology. Minneapolis, Minnesota: University of Minnesota Press, 1989.

Ottenbacher KJ, Tomchek SD. Measurement in rehabilitation research: Consistency versus consensus. Physical Medicine and Rehabilitation Clinics of North America 1993;4(3):463-73, Granger CV, Gresham GE, editors. New developments in functional assessment, vol. 4.

O'Connell J. Metrology: The creation of universality by the circulation of particulars. Social Studies of Science 1993;23:129-73.

Perline R, Wright BD, Wainer H. The Rasch model as additive conjoint measurement. Applied Psychological Measurement 1979;3(2):237-55.

Qayum M, Ortenberg K, Morstead R, Siddiqui F, Mipro RC Jr, Fisher WP Jr. Measuring functional status in rehabilitation: Louisiana Rehabilitation Institute vs. Uniform Data System [abstract]. LSU Department of Medicine, Section of Physical

Medicine & Rehabilitation, Residents' Research Day. New Orleans, LA: LSU School of Medicine, June, 1996.

Ramsay JO. Review of Foundations of Measurement, Vol. 1, by D. H. Krantz et al. Psychometrika 1975;40:257-62.

Rao N, Kilgore KM. Predicting return to work using a variety of assessment scales [abstract]. Archives of Physical Medicine and Rehabilitation 1990;71:763.

Rasch G. Probabilistic models for some intelligence and attainment tests. reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980. Copenhagen, Denmark: Danmarks Paedogogiske Institut, 1960.

Rasch G. On general laws and the meaning of measurement in psychology. In: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability. Berkeley, California: University of California Press, 1961: 321-33.

Rasch G. On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. Danish Yearbook of Philosophy 1977; 14:58-94.

Rating of hospitals is delayed in an effort for stronger data. New York Times 1993 June 23:A19.

Reckase MD. Adaptive testing: The evolution of a good idea. Educational Measurement: Issues and Practice 1989;8:3.

Schaffer S. Late Victorian metrology and its instrumentation: A manufactory of Ohms. In: Bud R, Cozzens SE, editors. Invisible connections: instruments, institutions, and science. Bellingham, WA: SPIE Optical Engineering Press, 1992: 23-56.

Shapin S. The invisible technician. American Scientist 1989 November-December;77:554-63.

Smith RM. Fit analysis in latent trait measurement models. Journal of Applied Measurement 2000; 1(2):199-218.

Spector W, Katz S, Murphy J, Fulton J. The hierarchical relationship between activities of daily living and instrumental activities of daily living. Journal of Chronic Disabilities 1987;40:481-9.

Stenner J, Smith III M. Testing construct theories. Perceptual and Motor Skills 1982; 55:415-26.

Stevens SS. On the theory of scales of measurement. Science 1946;103:677-80.

Subcommittee on Research Committee on Fundamental Science. Assessing fundamental science: a report from the Subcommittee on Research, Committee on Fundamental Science. Washington, DC: National Science and Technology Council, 1996.

Tatum DS. A measurement system for speech evaluation [dissertation]. Dissertation Abstracts International 1991;52(4):1301A. Chicago, Illinois: University of Chicago.

Thurstone LL. The measurement of values. Chicago: University of Chicago Press, Midway Reprint Series, 1959.

Thissen, D. MULTILOG, V. 6. Scientific Software, Inc.: Chicago, IL, 1991.

Wegener B. Outline of a structural taxonomy of sensory and social psychophysics. In: Wegener B, editor. Social attitudes and psychophysical measurement. Hillsdale, NJ: Lawrence Erlbaum, 1982: 1-42.

Weinsheimer J. Gadamer's hermeneutics: a reading of Truth and Method. New Haven, Connecticut: Yale University Press, 1985.

Weiss DJ. New horizons in testing: latent trait test theory and computerized adaptive testing. New York: Academic Press, 1983.

Weiss DJ, Kingsbury GG. Application of computerized adaptive testing to educational problems. Journal of Educational Measurement 1984;21(4):361-75.

Wernimont G. Ruggedness evaluation of test procedures. ASTM Standardization News 1977 March;5:13-6.

Wernimont G. Careful intralaboratory study must come first. ASTM Standardization News 1978 December;6:11-2.

Widmalm S. On exactitude, a review of The Values of Precision, by M. Norton Wise. Science 1995 May 12.

Wilson M. A comparison of deterministic and probabilistic approaches to learning structures. Australian Journal of Education 1989;33(2):127-40.

Woodcock RW. What can Rasch-based scores convey about a person's test performance? In: The new rules of measurement: what every psychologist and educator should know, edited by Embretson SE, Hershberger SL. Hillsdale, NJ: Lawrence Erlbaum Associates, 1999.

Wright BD. Sample-free test calibration and person measurement. In: Proceedings of the 1967 invitational conference on testing problems. Princeton, New Jersey: Educational Testing Service, 1968: 85-101.

Wright BD. Misunderstanding the Rasch model. Journal of Educational Measurement 1977;14(3):219-25.

Wright BD. Solving measurement problems with the Rasch model. Journal of Educational Measurement 1977;14(2):97-116.

Wright BD. Foreword, Afterword. In: Probabilistic models for some intelligence and attainment tests, by Georg Rasch. Chicago: University of Chicago Press, 1980.

Wright BD. Despair and hope for educational measurement. Contemporary Education Review 1984;3(1):281-8.

Wright BD. Additivity in psychological measurement. In: Roskam E, editor. Measurement and personality assessment. North Holland: Elsevier Science Ltd, 1985.

Wright BD, Bell SR. Item banks: What, why, how. Journal of Educational Measurement 1984;21(4):331-45.

Wright BD, Douglas GA. Best test design and self-tailored testing. Research Memorandum # 19. MESA Laboratory, Department of Education, University of Chicago, 1975.

Wright BD, Douglas G. The rating scale model for objective measurement. MESA Psychometric Laboratory Research Memoranda, vol. 35. Chicago: Dept of Education, University of Chicago, 1986.

Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Archives of Physical Medicine and Rehabilitation 1989;70(12):857-67.

Wright BD, Linacre JM. A users' guide to Bigsteps Rasch model computer program. Chicago: MESA Press, 1995.

Wright BD, Linacre JM, Heinemann AW. Measuring functional status in rehabilitation. Physical Medicine and Rehabilitation Clinics of North America 1993;4(3):475-91Granger CV, Gresham GE, editors. New developments in functional assessment, vol. 4.

Wright BD, Masters GN. Rating scale analysis: Rasch measurement. Chicago: MESA Press, 1982.

Wright BD, Mok M. Understanding Rasch measurement: Rasch models overview. Journal of Applied Measurement 2000; 1(1):83-106.

Wright BD, Stone MH. Best test design: Rasch measurement. Chicago: MESA Press, 1979.

Zhu W. Should total scores from a rating scale be used directly? Research Quarterly for Exercise and Sport 1996; 67(3):363-72.

Zhu W. Test equating: what, why, how? Research Quarterly for Exercise and Sport 1998; 69(1):11-23.

Zhu W, Updike WF, Lewandowski C. Post-hoc Rasch analysis of optimal categorization of an ordered response scale. Journal of Outcome Measurement 1997; 1(4):286-304.

Zupko RE. British weights and measures: a history from antiquity to the seventeenth century. Madison, WI: University of Wisconsin Press, 1977.

A3. SCIENTIFIC PUBLICATIONS DOCUMENTING RASCH-BASED INSTRUMENT CALIBRATIONS (NOT AN EXHAUSTIVE LIST)

A3.1 Comprehensive

A3.1.1 *Clinical*:

A3.1.1.1 *Patient Evaluation Conference System (PECS)*

Harvey RF, Fisher WP, Jr. The Patient Evaluation Conference System (PECS©). In: Collecting Information from Patients: A Resource Manual of Tested Questionnaires and Practical Advice, edited by McGee J, Goldfield N, Morton J, Riley K, 13:87-99. Gaithersburg, Maryland: Aspen Publications, Inc., 1997.

Harvey RF, Silverstein B, Venzon MA, Kilgore KM, Fisher WP Jr, Steiner M, et al. Applying psychometric criteria to functional assessment in medical rehabilitation: III. construct validity and predicting level of care. Archives of Physical Medicine and Rehabilitation 1992;73(10):887-92.

Kilgore KM, Fisher WP Jr, Silverstein B, Harley JP, Harvey RF. Application of Rasch analysis to the Patient Evaluation and Conference System. Physical Medicine and Rehabilitation Clinics of North America 1993;4(3):493-515.

Silverstein BJ, Fisher WP Jr, Kilgore KM, Harvey RF, Harley JP. Applying psychometric criteria to functional assessment in medical rehabilitation: II. defining interval measures. Archives of Physical Medicine and Rehabilitation 1992;73(6):507-18.

Silverstein BJ, Kilgore KM, Fisher WP Jr. Implementing patient tracking systems and using functional assessment scales. Center for Rehabilitation Outcome Analysis Monograph Series on Issues and Methods in Rehabilitation Outcome Analysis, vol. 1. Wheaton, Illinois: Marianjoy Rehabilitation Center, 1989.

A3.1.1.2 *Levels of Rehabilitation Scale—III (LORS)*

Velozo CA, Magalhaes LC, Pan A, Leiter P. Functional scale discrimination at admission and discharge: Rasch analysis of the Level of Rehabilitation Scale-III. Archives of Physical Medicine and Rehabilitation 1995;76(8):705-12.

A3.1.2 *Self-report*:

A3.1.2.1 *Louisiana State University Health Status Instruments (LSU HSI)*

Fisher WP Jr, Marier RL, Eubanks R, Hunter SM. The LSU Health Status Instruments (HSI). In: McGee J, Goldfield N, Morton J, Riley K, editors. Collecting Information from Patients: A Resource Manual of Tested Questionnaires and Practical Advice (Supplement). Gaithersburg, Maryland: Aspen Publications, Inc, 1997: forthcoming.

A3.2 Physical Function

A3.2.1 *Clinical*:

A3.2.1.1 *Functional Independence Measure (FIM)*

Chang W, Chan C, Slaughter S, Cartwright D. Evaluating the Phone-FIM Part II: Concurrent validity and influencing factors. Journal of Outcome Measurement 1997; 1(4):259-85.

Heinemann AW, Hamilton BB, Granger CV, Wright BD, Linacre JM, Betts HB, et al. Rating scale analysis of functional assessment measures. NIDRR Innovation Grant Award Final Report. Chicago, Illinois: Rehabilitation Institute of Chicago, 1991.

Heinemann AW, Linacre JM, Wright BD, Hamilton BB, Granger CV. Relationships between impairment and physical disability as measured by the Functional Independence Measure. Archives of Physical Medicine and Rehabilitation 1993;74(6):566-73.

Heinemann AW, Linacre JM, Wright BD, Hamilton BB, Granger CV. Prediction of rehabilitation outcomes with disability measures. Archives of Physical Medicine and Rehabilitation 1994;75(2):133-43.

Heinemann A, Segal M, Schall RR, Wright BD. Extending the range of the Functional Independence Measure with SF-36 items. In: Richard Smith RH Benjamin Wright, editor. Outcome Measurement in the Health Sciences, Vol. 1. First International Outcome Measurement Conference. Chicago: MESA Press, May 31–June 1, 1996: in press.

Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. Archives of Physical Medicine and Rehabilitation 1994;75(2):127-32.

Wright BD, Linacre JM, Heinemann AW. Measuring functional status in rehabilitation. Physical Medicine and Rehabilitation Clinics of North America 1993; 4(3):475-91.

A3.2.1.2 *Katz Activities of Daily Living Scale*

Teresi JA, Cross PS, Golden RR. Some applications of latent trait analysis to the measurement of ADL. Journal of Gerontology: Social Sciences 1989;44(5):S196-204.

A3.2.1.3 *Other Clinical*

Granger CV, Wright BD. Looking ahead to the use of functional assessment in ambulatory physiatric and primary care. Physical Medicine and Rehabilitation Clinics of North America: New Developments in Functional Assessment 1993; 4(3) August:595-605.

Zhu W, Cole EL. Many-faceted Rasch calibration of a gross-motor instrument. Research Quarterly for Exercise and Sport 1996; 67(1):24-34.

Zhu W, Kurz KA. Rasch partial credit analysis of gross motor competence. Perceptual and Motor Skills 1994; 79:947-61.

A3.3 Fine Motor Skills

A3.3.1 *Clinical*:

A3.3.1.1 *Assessment of Motor and Process Skills (AMPS)*

Bernspång B, Fisher AG. Differences between persons with right or left cerebral vascular accident on the Assessment of Motor and Process Skills. Archives of Physical Medicine and Rehabilitation 1995;76(12):1144-51.

Duran LJ, Fisher AG. Male and female performance on the Assessment of Motor and Process Skills. Archives of Physical Medicine and Rehabilitation 1996; 77(10) October:1019-24.

Fisher AG. The assessment of IADL motor skills: An application of many-faceted Rasch analysis. American Journal of Occupational Therapy 1993 April;47(4):319-29.

Fisher AG. Development of a functional assessment that adjusts ability measures for task simplicity and rater leniency. In: Wilson M, editor. Objective measurement: theory into practice. Vol II. Norwood, New Jersey: Ablex Publishing Corporation, 1994: 145-75.

A3.3.1.2 *Tufts Assessment of Motor and Process Skills (TAMP)*

Fisher AG. Commentary on Applicability of the hierarchical scales of the Tufts Assessment of Motor Performance for school-aged children and adults with disabilities by Haley and Ludlow. Physical Therapy 1992;72(3):202-4.

Haley SM, Ludlow LH. Applicability of the hierarchical scales of the Tufts Assessment of Motor Performance for school-aged children and adults with disabilities. Physical Therapy 1992;72(3):191-202.

Haley SM, Ludlow LH. Author response to Fisher's commentary on Applicability of the hierarchical scales of the Tufts Assessment of Motor Performance for school-aged children and adults with disabilities. Physical Therapy 1992;72(3):204-6.

Ludlow LH, Haley SM. Polytomous Rasch models for behavioral assessment: The Tufts Assessment of Motor Performance. In: Wilson M, editor. Objective measurement: theory into practice, Volume 1. Norwood, New Jersey: Ablex Publishing Corporation, 1992: 121-37.

Ludlow LH, Haley SM, Gans BM. A hierarchical model of a functional performance test in rehabilitation medicine: The Tufts Assessment of Motor Performance. Evaluation and the Health Professions 1992.

A3.4 Cognitive Function

A3.4.1 *Clinical*:

A3.4.1.1 *Functional Independence Measure (FIM)*

Heinemann AW, Hamilton BB, Granger CV, Wright BD, Linacre JM, Betts HB, et al. Rating scale analysis of functional assessment measures. NIDRR Innovation Grant Award Final Report. Chicago, Illinois: Rehabilitation Institute of Chicago, 1991.

Heinemann AW, Linacre JM, Wright BD, Hamilton BB, Granger CV. Relationships between impairment and physical disability as measured by the Functional Independence Measure. Archives of Physical Medicine and Rehabilitation 1993;74(6):566-73.

Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. Archives of Physical Medicine and Rehabilitation 1994;75(2):127-32.

A3.4.1.2 *Others*

Willmes K. Psychometric evaluation of neuropsychological test performance. In: von Steinbüchel N, von Cramon D, Pöppel E, editors. Neuropsychological rehabilitation. New York: Springer, 1992: 103-13.

A3.5 Health-Related Quality of Life

A3.5.1 *Self-report*:

A3.5.1.1 *General Health Questionnaire*

Andrich D, Van Schoubroeck L. The General Health Questionnaire: A psychometric analysis using latent trait theory. Psychological Medicine 1989; 19(2) May:469-85.

A3.5.1.2 *MOS SF-36*

Haley SM, McHorney CA, Ware JE Jr. Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. unidimensionality and reproducibility of the Rasch item scale. Journal of Clinical Epidemiology 1994;47(6):671-84.

McHorney CA, Haley SM, Ware JE. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. Journal of Clinical Epidemiology 1997; 50(4):451-61.

Raczek A, et al. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. International Quality of Life Assessment. Journal of Clinical Epidemiology 1998; 51(11) Nov:1203-14.

Segal M, Heinemann A, Schall RR, Wright BD. Extending the range of the Functional Independence Measure with SF-36 items. Physical Medicine & Rehabilitation: State of the Art Reviews 1997; 11(2) June:385-96.

Stucki G, Daltroy L, Katz N, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. Journal of Clinical Epidemiology 1996;49(7):711-7.

A3.6 Satisfaction with Services

A3.6.1 *Clinical:* None.

A3.7 Disease- and Population-specific Measures

A3.7.1 *Alcohol Abuse*:

Cornel M, Knibbe RA, van Zutphen WM, Drop MJ. Problem drinking in a general practice population: the construction of an interval scale for severity of problem drinking. Journal of Studies on Alcohol 1994; 55(4) July:466-70.

Gilbert FS. Development of a 'Steps Questionnaire.' Journal of Studies on Alcohol 1991; 52(4) July:353-60.

A3.7.2 *Arthritis*:

Nordenskiöld U, Grimby G, Hedberg M, Wright B, Linacre JM. The structure of an instrument for assessing the effects of assistive devices and altered working methods in women with rheumatoid arthritis. Arthritis Care and Research 1996; 9(5) October:358-67.

A3.7.3 *Asthma*:

Rosier MJ, Bishop J, Nolan T, Robertson CF, Carlin JB, Phelan PD. Measurement of functional severity of asthma in children. American Journal of Respiratory and Critical Care Medicine 1994; 149:1434-41.

A3.7.4 *Cancer*:

Zhu W, Smoczyk CM, Whatley M. An instrument for studying cancer patients' preferences for support groups. Journal of Cancer Education 1992; 7(3):267-79.

A3.7.5 *COPD*:

Prieto L, Alonso J, Ferrer M, Antó JM. Are results of the SF-36 Health Survey and the Nottingham Health Profile similar? A comparison in COPD patients. Journal of Clinical Epidemiology 1997; 50(4):463-73.

A3.7.6 *Diabetes*:

Lee NP, Fisher WPJ. Evaluation of the Diabetes Self Care Scale, in preparation, 2001.

A3.7.7 *Epilepsy*:

Gehlert S, Chang C. Factor structure and dimensionality of the multidimensional health locus of control scales in measuring adults with epilepsy. Journal of Outcome Measurement 1998; 2(3):173-90.

Sørensen AS, Hansen H, Andersen R, Høgenhaven H, Allerup P, Bolwig TG. Personality characteristics and epilepsy. Acta Psychiatrica Scandinavia 1989; 80:620-31.

A3.7.8 *Gerontology*:

Avlund K, Kreiner S, Schultz-Larsen K. Construct validation and the Rasch model: functional ability of healthy elderly people. Scandinavian Journal of Social Medicine 1993; 21(4) December:233-46.

Daltroy LH, Logigian M, Iversen MD, Liang MH. Does musculoskeletal function deteriorate in a predictable sequence in the elderly? Arthritis Care Research 1992; 5:146-50.

Doble SE, Fisher AG. The dimensionality and validity of the Older Americans Resources and Services (OARS) Activities of Daily Living (ADL) Scale. Journal of Outcome Measurement 1998; 2(1):4-24.

Lusardi MM, Smith EV. Development of a scale to assess concerns about falling and applications to treatment programs. Journal of Outcome Measurement 1997; 1(1):34-55.

Pollak N, Rheault W, Stoecker JL. Reliability and validity of the FIM for persons aged 80 years and above from a multilevel continuing care retirement community. Archives of Physical Medicine and Rehabilitation 1996; 77(10) October:1056-61.

A3.7.9 *Heart Disease*:

Siegrist J, Peter R, Junge A, Cremer P, Seidel D. Low status control, high effort at work, and ischemic heart disease: prospective evidence from blue-collar men. Social Science and Medicine 1990; 31(10):1127-34.

A3.7.10 *Pain*:

Kalinowski A. Measuring clinical pain. Journal of Psychopathology and Behavioral Assessment 1985;7:329-49.

McArthur DL, Casey KL, Morrow TJ, Cohen MJ, Schandler SL. Partial-credit modeling and response surface modeling of biobehavioral data. In: Objective measurement: theory into practice, vol. 1, edited by Wilson M, 109-20. Norwood, New Jersey: Ablex, 1992.

McArthur DL, Cohen M, Schandler S. Rasch analysis of functional assessment scales: an example using pain behaviors. Archives of Physical Medicine and Rehabilitation 1991; 72:296-304.

A3.7.11 *Pediatrics*:

Banerji M, Smith RM, Dedrick RF. Dimensionality of an early childhood scale using Rasch analysis and confirmatory factor analysis. Journal of Outcome Measurement 1997; 1(1):56-85.

Campbell SK, Kolobe THA, Osten ET, Lenke M, Girolami GL. Construct validity of the test of infant motor performance. Physical Therapy 1995; 75(7):585-96.

Haley SM, Ludlow LH, Coster WJ. Pediatric Evaluation of Disability Inventory: Clinical interpretation of summary scores using Rasch rating scale methodology. Physical Medicine and Rehabilitation Clinics of North America 1993;4(3):529-40.

A3.7.12 *Schizophrenia*:

Lewine RRJ, Fogg L, Meltzer HY. Assessment of negative and positive symptoms in schizophrenia. Schizophrenia Bulletin 1983; 9(3):368-76.

A3.7.13 *Etc*:

Gehlert S, Chang C, Hartlage S. Establishing the diagnostic validity of premenstrual dysphoric disorder using Rasch analysis. Journal of Outcome Measurement 1997; 1(1):2-18.

Grimby G, Andrén E, Holmgren E, Wright B, Linacre JM, Sundh V. Structure of a combination of Functional Independence Measure and Instrumental Activity Measure items in community-living persons: a study of individuals with spina bifida. Archives of Physical Medicine and Rehabilitation 1996; 77(11) November:1109-14.

Prieto L, Roset M, Badia X. Rasch measurement in the assessment of Growth Hormone Deficiency in adult patients. Journal of Applied Measurement 2001; 2(1):48-64.

Schulman JA, Wolfe EW. Development of a nutrition self-efficacy scale for prospective physicians. Journal of Applied Measurement 2000; 1(2):107-30.

Smith EV, Johnson BD. Attention Deficit Hyperactivity Disorder: Scaling and standard setting using Rasch measurement. Journal of Applied Measurement 2000; 1(1):3-24.

Tenenbaum G, Furst D, Weingarten G. A statistical reevaluation of the STAI anxiety questionnaire. Journal of Clinical Psychology 1985; 41(2):239-44.

Tesio L, Cantagallo A. The Functional Assessment Measure (FAM) in Closed Traumatic Brain Injury Outpatients: A Rasch based psychometric study. Journal of Outcome Measurement 1998; 2(2):79-96.

Whiteneck GG, Charlifue SW, Gerhart KA, Overholser JD, Richardson GN. Quantifying handicap: a new measure of long-term rehabilitation outcomes. Archives of Physical Medicine and Rehabilitation 1992; 73(6) June:519-26.