NOTICE: This standard has either been superseded and replaced by a new version or withdrawn.
Contact ASTM International (www.astm.org) for the latest information

Designation: E2171 – 02(Reapproved 2008)

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 E2171; 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 Practice E1384 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:*[2]

E177 Practice for Use of the Terms Precision and Bias in ASTM Test Methods

E456 Terminology Relating to Quality and Statistics

E691 Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method

E1169 Practice for Conducting Ruggedness Tests

E1384 Practice for Content and Structure of the Electronic Health Record (EHR)

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.

[1] This practice is under the jurisdiction of ASTM Committee E31 on Healthcare Informatics and is the direct responsibility of Subcommittee E31.25 on Healthcare Data Management, Security, Confidentiality, and Privacy.

Current edition approved Sept. 15, 2008. Published December 2008. Originally approved in 2002. Last previous edition approved in 2002 as E2171 – 02. DOI: 10.1520/E2171-02R08.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

Copyright © ASTM International, 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428-2959. United States

Documents
Document Sharing Hub

Full document details: ASTM E2171-02(2008)

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. ($P/1-P$).

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 Practice E1384 would necessarily have to stipulate fields for the EHR that

AVA Documents
Document Sharing Hub

Full document details: ASTM E2171-02(2008)

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 Practice E1384 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 Practice E1384 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)

AVA Documents
Document Sharing Hub

Full document details: ASTM E2171-02(2008)

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

AVA Documents
Document Sharing Hub

Full document details: ASTM E2171-02(2008)

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

AVA Documents
Document Sharing Hub

Full document details: ASTM E2171-02(2008)

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

AVA Documents
Document Sharing Hub

Full document details: ASTM E2171-02(2008)

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.[3] 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,

[3] 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 Terminology E456 and Practice E691 for further details on the contrast between repeatability and reproducibility.

AVA Documents
Document Sharing Hub

Full document details: ASTM E2171-02(2008)