Editorial

This issue of the Journal of Outcome Measurement (JOM) includes articles covering a variety of applications of measurement and Rasch analysis. Measurement in the psychosocial sciences is addressed by William P. Fisher, Jr., Ph.D. and he presents some thought provoking concepts and questions for the field of measurement. Your responses to these will be published and submitted to the author for comment. The remaining articles present a range of applications of Rasch measurement to physician work, cognitive growth and development, patterns of disability and ordinal scales, providing a superb overview of the power of measurement and in particular of the Rasch methodology.

Please consider the JOM for publication of your measurement and outcome studies.

Please call: (630) 221-1200 x 222
Or fax: (630) 221-1201
Or Email: rfharvey@rfi.org

Your Editor,

Richard F. Harvey, M.D.

JOURNAL OF OUTCOME MEASUREMENT®, 4(2), 527-563

Objectivity in Psychosocial Measurement: What, Why, How

William P. Fisher, Jr.
LSU Medical Health Sciences Center,
New Orleans, LA

This article raises and tries to answer questions concerning what objectivity in psychosocial measurement is, why it is important, and how it can be achieved. Following in the tradition of the Socratic art of maiuetics, objectivity is characterized by the separation of meaning from the geometric, metaphoric, or numeric figure carrying it, allowing an ideal and abstract entity to take on a life of its own. Examples of objective entities start from anything teachable and learnable, but for the purposes of measurement, the meter, gram, volt, and liter are paradigmatic because of their generalizability across observers, instruments, laboratories, samples, applications, etc. Objectivity is important because it is only through it that distinct conceptual entities are meaningfully distinguished. Seen from another angle, objectivity is important because it defines the conditions of the possibility of shared meaning and community. Full objectivity in psychosocial measurement can be achieved only by attending to both its methodological and its social aspects. The methodological aspect has recently achieved some notice in psychosocial measurement, especially in the form of Rasch's probabilistic conjoint models. Objectivity's social aspect has only recently been noticed by historians of science, and has not yet been systematically incorporated in any psychosocial science. An approach to achieving full objectivity in psychosocial measurement is adapted from the ASTM Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method (ASTM Committee E-11 on Statistical Methods, 1992).

Requests for reprints should be sent to William P. Fisher, Jr., Professor of Research, Public Health & Preventive Medicine, LSU Health Sciences Center, 1901 Perdido St., New Orleans, LA 70112-1393

> *"With the prospect of the development of a framework for merging scales ahead, the real problem is, of course, determining how the convergence of scales is to be accomplished, if, indeed, it can be accomplished at all." (Wegener, 1982, p. 41)*

Future generations will wonder how it ever came to pass that summed ratings were mistaken for measures. In contrast with most uses of the word "measure," the size and meaning of summed ratings' unit of measurement are dependent to unknown extents on the particular persons measured, the particular questions asked, the particular rating scale categories employed, and other factors, such as the particular clinician making the observations, or the facility where treatment is provided. Not only are these alleged measures scale-dependent and sample-dependent, they are commonly treated as interval or ratio measures, when they are not. They are ordinal, meaning that the data support only comparisons of rank order, not comparisons of amount.

Though they incorporate various unexamined dependencies and consist of a unit of measurement of unknown and changeable size, little, if anything, in the way of quality assessment and improvement is ever offered for psychosocial measures, largely because changes to the scale would also change the scale-dependent unit of measurement, making new data incommensurable with old. The inability of the method of summated ratings to deal with missing data is its primary practical flaw, compromising not only the commensurability of the scores, but the objectivity of comparisons based on them, the capacity to establish universal metrics referenced by all instruments measuring a particular variable, opportunities for removing variation in rater harshness or leniency from the measures, and the adaptive administration of scale items selected for their particular relevance to the individual measured.

As demands for accountability in education and health care continue to mount, more and more idiosyncratically and arbitrarily scored tests, health status, patient satisfaction, quality of life, and functional assessment instruments are added to the already cacophonous psychosocial measurement Tower of Babel. Recently, a number of articles have offered technical arguments for not treating ordinal data as interval, showing how the lack of a constant unit difference in the raw scores can lay the "foundations of misinference" (Merbitz, Morris and Grip, 1989; Fisher, 1993; Michell, 1990, 1997; Silverstein, Fisher, Kilgore, Harvey and Harley, 1992; Stucki, Daltroy, Katz, Johannesson and Liang, 1996; Wilson, 1971; Wright and Linacre, 1989; Zhu, 1996). Because of his concern that sociology is founded on unstable inferential bases, Wilson (1971, p. 433) strongly asserts that

> the task of developing valid, reliable interval measurement is not a technical detail that can be postponed indefinitely while the main efforts in sociological research are devoted to substantive theory construction; rather it is the central theoretical and methodological problem in scientifically oriented sociology.

This article is intended to show that the task of developing valid, reliable measurement does not stop at the point where ordinal data are transformed into scale- and sample-free interval measures in instrument calibration studies. Interval measures taken alone, isolated in different samples of persons and items in separate laboratories with no reference standards for metric range and unit or for data quality, do not solve the central theoretical and methodological problem of the psychosocial sciences.

The task of psychosocial measurement has another aspect that remains virtually unaddressed, and that is the social dimension of metrology, the networks of technicians and scientists who monitor the repeatability and reproducibility of measures across instruments, users, samples, laboratories, applications, etc. For the problem of valid, reliable interval measurement to be solved, within-laboratory results must be shared and communicated between laboratories, with the aim of coining a common currency for the exchange of quantitative value. Instrument calibration (intralaboratory repeatability or ruggedness) studies and metrological (interlaboratory reproducibility) studies must be integrated in a systematic approach to accomplishing the task of developing valid, reliable interval measurement. Accordingly, this article will first pose the question concerning what objectivity in psychosocial measurement is; then it will address the questions of why it is important, and how it can be achieved.

What is objectivity in psychosocial measurement?

The popular distinction between subjective and objective measures (McDowell and Newell, 1996, p. 14) is more like a bad novel than anything that has to do with the history of science and reason (Heidegger, 1967, p. 99; Lakoff and Johnson, 1980). Common usage in health care outcome measurement holds that self-reported satisfaction, attitude, or health status measures are subjective, and blood pressure, body temperature, height, and weight are objective. This sense of the subjective as allegedly unobservable,

internal, or mental, and of the objective as allegedly observable, concrete, and material, has nothing to do with the philosophical senses of subjectivity and objectivity that have made science possible. Objectivity is not concreteness, and subjectivity is not the idiosyncratic experience of an individual.

Methodologically, the problem of objectivity in psychosocial measurement has much in common with the problem of objectivity in geometrical measurement faced by Plato 2500 years ago. Plato drew from his teacher Socrates' sense of objectivity as a matter of maiuetics (midwifery). As is recounted in the *Theaetetus* (149a-151e), Socrates saw his task as one of aiding in the birth of ideas, and checking them to see whether "the offspring of ... thought is a false phantom or instinct with life and truth" (150c). Plato applied Socrates' method to the problem of line segments of irrational length, as the hypotenuse of a right isosceles triangle must be. Plato realized that irrational numbers are not false phantoms simply because they represent line segments that cannot be made commensurable with other line lengths, no matter how precisely they are drawn, as was held by the Sophists and Pythagoreans. Plato showed that irrational numbers have just as much objective existence and mathematical validity as rational numbers insofar as they come alive with meaning in the context of geometry. Plato then restricted the instruments of geometry to the compass and the straightedge because only these, and not the mechanical devices employed by Sophists and Pythagoreans for copying and manipulating figures, allowed the meaningful, abstract ideality conveyed by the figures to separate from them and take on lives of their own (Fisher, 1992).

All measurement must therefore be recognized as indirect, since abstract ideas are never observable in and of themselves. It has long been recognized that the objectivity of weight, for instance, is "customarily determined by watching a pointer on a scale. No one could truthfully say that he 'saw' the weight" (Guilford, 1936, pp. 1-19; Andrich, 1990; Rasch, 1977, p. 68; Rehfeldt, 1990, p. 117). Even perception itself, because it is a process of selectively focussing attention, is a *reading*, and hence interpretation, of the world, making all of the implications of hermeneutics (interpretation theory) relevant to science (Heelan, 1972, 1983a, 1983b, 1983c; Ihde, 1991; Stent, 1981; Nicholson, 1984; Fisher, 1988, 1991, 1992; Weinsheimer, 1985).

Objectivity is not a special epistemological advantage that physical measures have over measures of human abilities, attitudes, or health. Nonetheless, data from rating scales filled out by clinicians are typically considered more objective than self-reported data because observations are tied to concrete clinical indicators, such as the percent to which someone with disabilities is functionally dependent on a caregiver.

Following Socrates and Plato, then, we see that objectivity has a methodological aspect characterized by the abstract and ideal stability, repeatability, and reproducibility of a unit of measurement within and across samples, instruments, laboratories, operators, environmental conditions, etc. But objectivity does not happen by itself or exist in nature apart from human culture and technology. It emerges only from within particular kinds of Socratic conversations, conversations that can be deliberately crafted by those who value them. Perhaps because of the transparency of instruments (Bud and Cozzens, 1992), the invisibility of technicians (Shapin, 1989), and facile assumptions about the mathematical relation of numbers to the things they represent (Duncan, 1984; Fisher, 1992; Merbitz, et al., 1989; Michell, 1990, 1997; Wilson, 1971; Wright and Linacre, 1989; Zhu, 1996), the psychosocial sciences lack the social networks of technicians who focus on converting experiments into governable instruments, and then on maintaining and improving the stability of the instruments' units of measurement across all of the possible sources of variation. The social and methodological senses of objectivity will be explored in more detail before taking up the why and the how of objectivity.

Social Objectivity

Without embarking on a long excursion into the large literature on the topic (Ackermann, 1985; Bernstein, 1983; Brown, 1977; Bud and Cozzens, 1992; Daston, 1992; Daston and Galison, 1992; Hacking, 1983; Heelan, 1965, 1983a, 1983b, 1993; Hesse, 1972; Ihde, 1991; Kuhn, 1970; Latour, 1989; Mendelsohn, 1992; Schaffer, 1992; Shapin, 1989; Todes and Dreyfus, 1970; Wise, 1988), it must be pointed out that both of objectivity's primary aspects, the social and methodological, must be incorporated into any science that hopes to be quantitative. Although in current Rasch measurement practice objectivity is defined in purely methodological terms, success in quantification cannot be restricted to the use of well-crafted instruments that meet specific mathematical requirements for parameter invariance (see next section).

In addition to methodological objectivity, quantitative success must also require groups of people who 1) agree on methods for assessing data quality, 2) agree on quantitative unit sizes and ranges, 3) agree on the skills required for instrument use, and who 4) circulate reference standard samples

and instruments among themselves, maintaining and enhancing the standards according to longstanding and widely accepted methods of conducting interlaboratory consistency trials (ASTM Committee E-11 on Statistical Methods, 1992; Mandel, 1977; Mandel, 1978; Wernimont, 1977; Wernimont, 1978). As O'Connell (O'Connell, 1993, p. 166) puts it,

> The challenge to social scientists is to stop taking the universality of entities like the volt for granted, and to start treating the volt like the *society* which it truly is—a distributed collective connected by continually renewed structured relations of exchange and authority. Scientific entities are not universal until scientists or their technicians take the trouble to make them so.

Similarly, Widmalm (Widmalm, 1995) says that

> Technical reliability depends on the organization of people. Scientific conventions, such as standards, are both agents of unity and products of agreement. . . . The adoption of standards reflects the adoption of laboratory cultures. Precision is the result not only of individual technical prowess, but of networks of scientists who rely on an infrastructure of workshops and bureaucracies.

And Schaffer (Schaffer, 1992) says that

> The physical values which the laboratory fixes are sustained by the social values which the laboratory inculcates. . . . In milieux such as those of Victorian Britain the propagation of standards and values was the means through which physicists reckoned they could link their work with technical and economic projects elsewhere in their society. Instrumental ensembles let these workers embody the values which mattered to their culture in their laboratory routines. Intellectualist condescension distracts our attention from these everyday practices, from their technical staff, and from the work which makes results count outside laboratory walls.

Even in the field of historical metrology, "weights and measures officials are almost never discussed, and if they are they receive only brief mention as a footnote to a presentation of the units" (Zupko, 1977, p. xiv). Thus, the social sciences do not have objective units of measurement that are universally recognized and accepted at least in part because they do not take the trouble to circulate and evaluate well-calibrated instruments and samples of known value for conformity with the chosen standards.

As social studies of science and technology continue to focus on the roles played by instruments and technicians in sustaining the physical sciences' universal units of measurement, a context and need will be created for the emergence of methods that test numbers produced by rating scale instruments for the mathematical invariance required for establishing universal units of measurement in the psychosocial sciences. Methods already in place (ASTM Committee E-11 on Statistical Methods, 1992; Mandel, 1977; Mandel, 1978; Wernimont, 1977; Wernimont, 1978) for evaluating the consistency of physical measures and calibrations within and across laboratories, samples, operators, etc. provide a model for how to create, maintain, and improve universal metrics for specific variables accessed via rating scales (Fisher, 1997b). In addition to these methods, and in addition to the technicians interested in fostering networks of social objectivity, the psychosocial fields employing rating scale instruments must also pay strict attention to a more fundamental level of methodological objectivity.

Methodological Objectivity

L. L. Thurstone's (Thurstone, 1959, p. 228) 1928 definition of objectivity in measurement is as relevant now as ever:

> One crucial experimental test must be applied to our method of measuring attitudes before it can be accepted as valid. A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument is impaired or limited. If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement.... If the scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help to construct it. This may turn out to be a severe test in practice, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale.

Thurstone's method of paired comparisons came very close to providing the tests needed to establish whether or not scale-free person measurement, and sample-free instrument calibration, had been achieved. Other ways of stating Thurstone's requirement for scale-free measurement were offered by Jane Loevinger (Loevinger, 1947) and Louis Guttman (Guttman, 1950). Guttman says that the "definition of a scale ... requires that each person's responses

should be reproducible from his rank alone," "rank" referring to a position in a distribution defined by the person's summated ratings, or score. For this reproducibility to occur, the order of the items on the instrument cannot be dependent on the particular person measured, just as Thurstone required.

The works of Thurstone and Guttman have become part of a tradition in measurement theory that continues to the present (Bradley and Terry, 1952; Cliff, 1973; Krantz, Luce, Suppes and Tversky, 1971; Luce and Tukey, 1964; Lumsden, 1980; Michell, 1990; Perline, Wright and Wainer, 1979; Wright, 1997). The admitted and recognized problem with this line of theory development is its requirement of an absolute, deterministic order to survey questions. As Guttman (Guttman, 1950, p. 63) says,

> Murphy, Murphy, and Newcomb further note that, 'As a matter of fact there is every reason to believe that none of the rather complex social attitudes which we are primarily discussing will ever conform to such rigorous measurement.' Perhaps such a belief may account for the fact that the mass of current attitude research pays little or no attention to this fundamental rationale. The common tendency has been to plunge into analysis of data without having a clear idea as to when a single dimension exists and when it does not. For example, *bi*variate techniques—like critical ratios and biserial correlations—are commonly used to find items that 'discriminate' and to determine 'weights,' without testing whether or not the *multi*variate distribution of the items is actually indicative of a single dimension.

This comment dates from 1950, but could just as easily be applied to the vast majority of today's psychosocial measurement applications. What Guttman calls plunging into data analysis without checking to see whether a single dimension exists to support the meaning of a summated score is still far more common than not. Even when some concern is paid to dimensionality, it is usually approached by means of factor analysis, which does nothing to test for or establish scale-free measurement.[1]

Applications of Thurstone's method of paired comparison are today practically nonexistent, and Guttman scaling often requires rejection of more than half of the data gathered as unscalable, so it too is rarely used. In 1946, Gulliksen (Gulliksen, 1946) unsuccessfully tried to revive the paired comparisons method and bring it to the attention of his colleagues. Cliff's (Cliff, 1973) review of advances made during the 1960s in measurement theory considered a paper by Luce and Tukey (Luce and Tukey, 1964) that used a

Thurstonian approach in reinventing fundamental measurement one of the two most seminal works of the period. Cliff had to admit his "disappointment that more advantage has not been taken of the [opportunities offered by the theory]."

In 1990, Michell (Michell, 1990, p. 67) essentially repeated Cliff's remark, contending that psychologists "have failed to realize the significance of Luce and Tukey's (1964) development" of "a new kind of fundamental measurement." Echoing Guttman, Michell (Michell, 1990, p. 130) says,

> In general psychologists have ... found refuge in quantitative methods that, because they assume more, demand less foundational research as the basis for their application. Methods that always yield a scaling solution, like the method of summated ratings, are almost universally preferred to methods which ... do not produce a scaling solution when they are falsified by the data. Surprisingly, vulnerability to falsification is commonly deemed by psychologists to be a fault rather than a virtue.

Part of the problem in convincing psychologists and others to treat the quantitative status of a variable as a falsifiable hypothesis has been the complexity of the theories and methods available for testing such hypotheses and creating data amenable to such tests, besides the unavailability of simple and easy-to-use software for these kinds of data analysis.

In contrast with the problems associated with deterministic approaches to measurement, probabilistic models for measurement worked out by Georg Rasch (Rasch, 1960; Rasch, 1961; Rasch, 1977) and built on by his students and colleagues (Andersen, 1980; Andrich, 1978a; Andrich, 1988; Fischer, 1973; Fischer, 1974; Fischer and Molenaar, 1995; Wright, 1968, 1977b; Wright and Masters, 1982; Wright and Stone, 1979) are much simpler and applicable to many kinds of data, with software (Adams and Kboo, 1995; Allerup and Sorber, 1977; ASC, 1996; Andrich, 1997; Andrich, Lynne and Sheridan, 1990; Glas and Ellis, 1995; Gustafsson, 1979; Kelderman and Steen, 1988; Linacre, 1995, 1998; Smith, 1991b; Wright and Linacre, 1995; Verhelst, 1993; Wu, Adams and Wilson, 1996) that has been undergoing continuous improvement for over 30 years. Relationships between Rasch's models and Guttman scaling, Thurstone's method of paired comparisons, Luce and Tukey's additive conjoint measurement principles, and Fisher's notion of sufficiency are well-known (Andersen, 1977; Andrich, 1978b; Andrich, 1985; Andrich, 1988; Brink, 1972; Brogden, 1977; Englehard, 1984; Englehard, 1994; Fisher and Wright, 1994; Green, 1986; Keats, 1967; Perline, et al., 1979; van der Linden, 1994; Wilson, 1989; Wright, 1980, 1997).

Rasch's entry into measurement theory hinged on the use he was able to make of Fisher's (Fisher, 1922) notion of statistical sufficiency, providing, in effect, a formal mathematical basis for Thurstone's "crucial experimental test" and for Guttman's sense of reproducibility, though Rasch was apparently unfamiliar with Thurstone's and Guttman's works at the time. Rasch was primarily a mathematician, but had worked with Fisher in London in the 1930s and was impressed by the mathematical implications of sufficiency. When Rasch was asked to help solve some educational measurement problems in Denmark in the 1950s, he made a point of making sure that the measures were statistically sufficient estimates of ability. He focussed on sufficiency because, as he (Rasch, personal communication recorded in Wright, 1980: xii) later said,

> When a sufficient estimate exists, it extracts every bit of knowledge about a specified feature of the situation made available by the data as formalized by the chosen model. 'Sufficient' stands for 'exhaustive' as regards the feature in question.
>
> What is left over when a sufficient estimate has been extracted from the data is independent of the trait in question and may therefore be used for a control of the model that does not depend on how the actual estimates happen to reproduce the original data....
>
> The realization of the concept of sufficiency, I think, is a substantial contribution to the theory of knowledge and the high mark of what Fisher did.... His formalization of sufficiency nails down the ... conditions that a model must fulfill in order for it to yield an objective basis for inference.

Sufficiency can be illustrated via the simple example of any common, everyday measurement task. Whenever a ruler, bathroom scale, or thermometer is read, what one is in effect doing is asking, "Is the indicator at or above here?" for every hash mark on the number line. Responses of "Yes" to this question are assigned a score of "1" and responses of "No" a score of "0". The measure is the sum of these scores. *The fundamental and inescapable requirement of measurement is that the pattern of "Yes" and "No" responses on the instrument, and on any other instrument measuring the same variable, be reproducible from the score alone.* Summary scores capable of meeting this requirement are sufficient statistics.

Instead of treating sufficiency as an assumption that can be ignored, Rasch turns it into a requirement that enables checks on the extent to which the requirements for measurement have been met in every response (Adams and Wright, 1994; Gustafsson, 1980; Ludlow, 1985; Smith, 1986, 1991a, 1994; Wright, 1977b, 1985, 1997; Wright and Masters, 1982; Wright and Stone, 1979). Control of the model follows from the fact that scores imply specific patterns of response if they are sufficient statistics, with the easiest tasks most likely to be succeeded on, and the most difficult ones, least likely. When data contradict the model, it is not the model that is at fault, if we value sufficiency, generalizability, objectivity, and the opportunity for creating universal metrics, but the data. For instance, ambiguously worded questions may provoke one response from some people, and another response from others, quite apart from the consistency of their responses on other questions. Similarly, questions or tasks that are much too easy or too difficult for an individual may provoke unexpected responses through carelessness, guessing, or special strengths and weaknesses; adaptively restricting questions to those of a difficulty close to the person's ability eliminates such problems.

Any use of summed ratings as measures assumes not only that these ordinal data are interval measures, but that each score is a sufficient statistic (Andersen, 1977). Rasch deliberately required sufficiency when he theorized that 1) a person with a higher score than another (assuming a common set of items) should also have a higher probability of success on any item, and 2) an item with a higher score than another (assuming administration to a common sample of persons) should also have a higher probability of being succeeded on by any person.[2] This is nothing but a formalization of what is necessarily assumed whenever scores are treated as measures. Thus,

> for anyone who claims scepticism about 'the assumptions' of the Rasch model, those who use unweighted scores are, however unwittingly, counting on the Rasch model to see them through. Whether this is useful in practice is a question not for more theorizing, but for empirical study (Wright, 1977b, p. 114).

Each new application of a Rasch model poses the question of whether counts of correct answers or of rating scale steps comprise useful and manageable frameworks for observing and measuring amounts of ability, attitude, or health. Eleven separate analyses of four different physical disability measures not only provide individual confirmations of the practical utility of counting on the Rasch model, but they provide strong evidence in support of the hypothesis that they measure the same variable, and could do so in the same metric (Fisher, 1997a).

Because of widespread misunderstanding of what Rasch measurement

is, it is necessary to point out that the area of theoretical analysis relevant to discussion of Rasch's models for measurement is fundamental measurement theory and related principles of mathematical invariance involving statistical consistency and sufficiency (Andrich, 1988; Fisher, 1994; van der Linden, 1994; Wright, 1977a, 1984, 1997). This point has been insufficiently appreciated in some summaries of the health measurement literature, where Rasch measurement mistakenly has been situated in the context of Item Response Theory (IRT). Neither of two recent books (Streiner and Norman, 1995; McDowell and Newell, 1996) makes any mention of Thurstone's (Thurstone, 1959) crucial experimental test, Fisher's (Fisher, 1922) sense of sufficiency, or fundamental measurement theory (Campbell, 1920; Luce and Tukey, 1964; Krantz, et al., 1971; Michell, 1990, 1997), all of which are commonly related to Rasch measurement (Andrich, 1978b; Andrich, 1988; Brogden, 1977; Englehard, 1984, 1994; Fisher and Wright, 1994; Perline, et al., 1979; van der Linden, 1994; Wright, 1980, 1977a; 1984, 1997; Wright and Masters, 1982; Wright and Stone, 1979). All of the reasons listed by Streiner and Norman (Streiner and Norman, 1995, p. 187) as to why Rasch measurement has not seen more application are mistaken. In fact, their claim that "latent-trait theory has not been widely used in test construction" contradicts the fact that Rasch measurement has been used for as long as 30 years by researchers at American College Testing, Educational Testing Service, the Psychological Corporation, and American Guidance Services; by educational research groups in Australia, Britain, Denmark, Israel, Malaysia, the Netherlands, South Africa, and elsewhere; and by school boards in Chicago, New York, Phoenix, Portland (OR), Minneapolis, and elsewhere.

As a result of the confusion about measurement theory, Streiner and Norman (1995, pp. 186-7) mistakenly attribute, or at least seem to attribute, Rasch measurement's advantages, such as scale-free measurement properties, to IRT, and IRT's shortcomings, such as the need for large sample sizes, to Rasch measurement. IRT models are not theories of scale-free measurement because they incorporate parameters, intended to describe item functioning, that "destroy the possibility of explicit invariance of the estimates" (Andrich, 1988, p. 67; Wright, 1977a; Wright, 1984; Fisher, 1994), meaning that scale-dependencies can and do remain in data fitting IRT models with 2 or 3 item parameters.

Regarding sample size, even one of the founders of IRT, Frederic Lord (Lord, 1983) recognized and accepted the value of Rasch models in the analysis of small data sets, since large samples are often required for multiparameter IRT models' estimates to converge. Streiner and Norman also are mistaken in claiming that data from "chained" questions, ones in which responses to different questions are interdependent, cannot be modelled to provide sufficient statistics because this kind of data violate the "assumption" of local independence. On the contrary, there is considerable interest among Rasch theoreticians and practitioners in just these models (Linacre, 1991a; Linacre, 1991b; Verhelst and Glas, 1993).

In accord with the IRT perspective, both Streiner and Norman (1995, p. 187), and McDowell and Newell (1996, p. 22), incorrectly view the need for experimental tests of parameter invariance as an inconvenient limitation. Both books try to justify abandoning Rasch's prescriptive measurement requirements in favor of other, less restrictive models that better describe the data, when data do not provide sufficient statistics and so do not fit the model specified.

IRT models parameters that abandon sufficiency and the *prescription* of measurement requirements, in favor of improved *description* of the data at hand. O. D. Duncan (Duncan, 1984, p. 217), a sociologist and social measurement expert with a long career in survey research, objects to this approach, saying

> In my view, what we need are not so much a repertoire of more flexible models for describing extant tests and scales ... but scales built to have the measurement properties we must demand if we take 'measurement' seriously. As I see it, a measurement model worthy of the name must make explicit some conceptualization--at least a rudimentary one--of what goes on when an examinee solves test problems or a respondent answers opinion questions; and it must incorporate a rigorous argument about what it *means* to measure an ability or attitude with a collection of discrete and somewhat heterogenous items.... Thurstone explicated the meaning of measurement as it might accomplished by such an instrument. Rasch provided the formalization of that meaning.

The reasons for this contrast between IRT's focus on data description and measurement theory's concern with data prescription are probably rooted in the value that educational measurement has historically placed on content validity (hence IRT's fitting of models to data), which is not shared by the mathematician's (Rasch) and the physicist's (Wright) valuation of construct validity (fitting data to models) (Fisher, 1994). Paraphrasing Michell (Michell, 1990, p. 130), IRT models are another example of the general

dislike in the psychosocial sciences for scaling methods that do not provide a measurement solution when they are falsified by the data. Hopefully, the growing interest in Rasch measurement is a sign that vulnerability to falsification may soon be deemed more of a virtue than a fault.

Why is objectivity in psychosocial measurement important?

The scientific advantages of objective measurement include data quality assessment and improvement, scale-free and sample-free units of measurement, and universal metrics. Alongside the scientific advantages associated with objective measures are several practical advantages. These follow primarily from the capacity of Rasch's models to take missing data into account. Probabilistically modelled scale-free measurement standards make it possible for every health measurement information need to be met with numbers that mean the same thing, within the error of measurement, no matter which brand instrument the measures come from (Cella, Lloyd and Wright, 1996; Fisher, Harvey, Taylor, Kilgore and Kelly, 1995; Fisher, Harvey and Kilgore, 1995; Fisher, Eubanks, and Marier, 1997; Fisher, 1997a; Gonin, Lloyd and Cella, 1996; Grimby, et al., 1996; Segal, Heinemann, Schall and Wright, 1997; Tennant and Young, 1997), and no matter which particular collection of calibrated items is administered (Choppin, 1976; Fisher, Harvey, Taylor, et al., 1995; Lunz, Bergstrom and Gershon, 1994; Kelderman, 1986; Masters, 1985; Smith and Kramer, 1992; Wright and Bell, 1984). No technical barriers prevent users from adapting instruments to their needs in order to obtain with the greatest efficiency the most reliable and relevant information possible. Instead of adapting our needs to the demands of the measurement technology, it is now possible to adapt the measurement technology to our needs.

In adaptive measurement, a survey respondent, examinee, clinical observer, examiner, or computer selects items for administration based on information concerning the intended application, special needs of the respondent, or the most efficient targeting of the instrument (Choppin, 1976; Esdaille, Shaw, Smith and Valgeirsdóttir, 1994; Lunz, et al., 1994; Reckase, 1989; Smith, et al., 1994; Weiss, 1983; Weiss and Kingsbury, 1984; Wright and Bell, 1984; Wright and Douglas, 1975). Given a set of calibrated items presented in measure order, respondents or raters could be instructed to provide at least some minimum number of ratings, with the understanding that items irrelevant to the respondent's lifestyle, culture, or treatment needs could be skipped, as could items representing tasks far too easy or difficult for the person to perform. This procedure can be implemented using either paper-and-pencil or computerized instruments, but it appears that the computational advantages of networked, handheld devices will be crucial to making adaptive measurement widespread.

There is another advantage of the use of calibrated item banks, one related to the way in which the embodiment of crucial experiments in instruments has been crucial to the creation of consistently manageable and observable phenomena in the history of science (Bud and Cozzens, 1992; Ihde, 1991; Price, 1986). Popular perceptions of the relation of science to technology assume that technology is in some way a product of science. Examination of the details of scientific practice reveals that the opposite is more usually the case.

For instance, "thermodynamics owes much more to the steam engine than ever the steam engine owed to thermodynamics," and "the chemical revolution resulted much more from the technique of the electric battery than from the careful measurements or new theories of Lavoisier" (Price, 1986, pp. 240, 248). In the same way, the widespread availability of standard health status constants, embodied in universally-accessible computerized banks of calibrated items and expressed in scale-free, sample-free, and variable-specific common metrics, could provide a focus in the theory and practice of health care that could lead to significant new public health advances, especially in the area of lifestyle-based preventive care.

Because each of the many diverse groups using functional assessment data (Costich, 1993) understandably want instruments that provide the level of detail they find most convenient for their purposes, health care has become overpopulated with functional status and health status scales, each of which measures in its own unit. This situation will need to change as medical records become, first, computerized on centralized, hospital-specific systems, and then transferred to decentralized, virtual patient records accessible from any terminal in the world with a link to the global network; information will have to be communicated via standardized content and structure if it is to be understood by all potential users (Board of Directors of the American Medical Informatics Association, 1994; ASTM Committee E-31 on Computerized Systems, 1996; Fisher, 1996, 1997c). If functional status measures are to play any role in this context, it will be only insofar as general, scale-free units of measurement can be brought to bear. If the current array of incommensurable health status measures were placed in that context, they would probably cease to be of any use, and could possibly become barriers to effective health care. The many existing successful applications of objective psychosocial measurement's item banking and instrument equating methods present a hopeful future for universal metrics of health status

measurement.

Objective psychosocial measurement data quality standards also make it possible to evaluate instruments in terms of the consistency of the data they produce, the extent to which they are relevant to the population of interest, and the range of error in which their measures must be interpreted. If existing instruments produce relatively inconsistent ratings in one or another area, new items can be developed to replace those introducing the inconsistencies without making the old data incommensurate with the new (Holm and Kavanagh, 1985; Wright, 1993; Wright and Stone, 1979). If a new instrument extends the range of measurement to include persons with more extreme abilities or attitudes, the value of that instrument is made far more evident by equating it with existing instruments than by correlating the old and new instruments' raw scores. In fact, if the old instrument is seriously off-target, very low correlations might result, and the new instrument might be mistakenly judged to measure a different variable than the old. A new instrument might increase the specificity of observations and lower measurement error accordingly.

One of the most recently developed practical advantages of Rasch measurement concerns multifaceted models that adjust measures for variations in raters' and judges' propensity to assign harsh or lenient scores (Linacre, 1989; Linacre, 1995). It has long been recognized that judges can introduce as much variation into examinees' scores as exists in the differences among the examinees' abilities (Cason and Cason, 1984; Edgeworth, 1890; Linacre, 1989, pp. 6, 21; Ruch, 1929; Ruggles, 1911). Even in the best of circumstances, with well-trained raters and a carefully designed instrument, agreement among judges on ratings is far from perfect (Borman, 1978; Linacre, 1989, p. 21).

Despite the lack of agreement, judges are often remarkably consistent in their ordering of item difficulties and person abilities (Lunz, Stahl and Wright, 1996), making it possible to include variation in judge rating behavior as a parameter in a Rasch model. Multifaceted models have been successfully applied to measurement problems in several fields (Linacre, Englehard, Tatum and Myford, 1994), including professional certification (Lunz, Wright and Linacre, 1990; Lunz, et al., 1994; Lunz, et al., 1996; Lunz and Stahl, 1993a; Lunz and Stahl, 1993b; Stahl and Lunz, 1996), occupational therapy (Fisher AG, 1993, 1994b; Fisher AG, et al., 1994a; Fisher and Fisher, 1993), sports performance (Fisher PB, 1993; Fisher PB, 1995); public speaking (Tatum, 1991), aesthetic judgment (Myford, 1989), medical clerkship evaluations (Fisher, Vial, and Sanders, 1997), and education (Hess, 1995; Englehard, 1992; Myford, Marr, and Linacre, 1996; Myford and Mislevy, 1995). Multifaceted Rasch models make it possible to assess and improve the quality of data from any number of identifiable but uncontrollable factors that consistently influence the outcome of the measurement process. Experimental designs incorporating these models can be expected to more clearly identify and remove sources of unwanted variation, thereby better isolating and estimating treatment effect sizes.

How is objectivity in psychosocial measurement attained?

Standards organizations such as ASTM identify two basic phases in the calibration of objective measurement systems (ASTM Committee E-11 on Statistical Methods, 1992; Mandel, 1977; Mandel, 1978; Wernimont, 1977; Wernimont, 1978), a practice known as metrology. The first of these phases is intralaboratory calibration, corresponding to the methodological level of objectivity described above, which tests the hypothesis that a single mathematically invariant variable is measured by an instrument. The extensive literature on Rasch measurement is focussed primarily on the intralaboratory calibration of tests and rating scales. Accessible introductory texts on Rasch measurement (Wright and Masters, 1982; Wright and Stone, 1979) can be combined with software (Wright and Linacre, 1995) and thoroughly illustrated examples that make it easy for beginners to get started. A promising new offering in this regard is Bond and Fox's (2001) new book.

Attention is only just beginning to be focussed on the second phase in the calibration of objective measurement systems, involving the social aspect of objectivity. The possibility that the intralaboratory phase might not be the whole story of objective measurement began to arise when a common construct was noticed in separate calibrations of the physical disability subscales of the Patient Evaluation Conference System (Harvey and Jellinek, 1981; Silverstein, et al., 1989; Silverstein, et al. 1992; Kilgore, et al., 1993) and of the Functional Independence Measure (Hamilton, Granger, Sherwin, et al., 1987; Heinemann, et al., 1991; Heinemann, Linacre, Wright, Hamilton and Granger, 1993; Heinemann, Linacre, Wright, Hamilton and Granger, 1994; Linacre, Heinemann, Wright, Granger and Hamilton, 1994; Wright, Linacre and Heinemann, 1993). Subsequent common sample equating of the two instruments showed that they could measure in a common metric (Fisher, et al., 1995; Smith, 1998). Further research examining the variable structures of four instruments calibrated in eleven separately-conducted studies produced an average correlation of .93 among the item calibrations (Fisher, 1997a). A trial application of ASTM interlaboratory

precision assessment methods on a sample of these data (Fisher, 1997b) provides further confidence in the likelihood of creating universal metrics for the measurement of physical disability, and beyond that variable to others measured using rating scale instruments.

What sufficiency looks like

Although a complete implementation of a Rasch analysis cannot be developed here, some of the basic concepts can be touched upon. Table 1 presents some hypothetical data organized in Guttman's scalogram pattern, with the data ordered vertically by the respondents' or examinees' scores, and horizontally by the items' scores. Because these data result in a pattern of overlapping triangles of 1s and 0s, they are conjointly ordered. The most difficult item, with the lowest score, is the least likely to be succeeded on for any person, no matter what their ability. Similarly, the easiest item, with the highest score, is the one most likely to be succeeded on by any person. Conversely, the person with the lowest score is the one least likely to succeed on any item, and the person with the highest score, most likely. Both the person and item scores are sufficient statistics since they allow reproduction of their associated patterns of responses to within an error of measurement (not estimated here).

The natural logarithm of the odds of success for the persons, and of failure for the items, as shown in Table 1, expresses the measures and calibrations in a common unit of measurement. This makes it possible to interpret a measure in terms of the likelihood of success on an item. For instance, a person with a measure of 0.00 logits, in the middle of Table 1's measurement continuum, has about a 50-50 chance of success on the three items in the middle of the scale, but a greater chance of success on the easier items (those with calibrations lower than 0.00), and a lesser chance of success on the more difficult items (those with calibrations higher than 0.00).

In addition to logit estimates of person ability and item difficulty, most Rasch measurement software provides error estimates and model fit (statistical consistency, or data quality) indices for each person and item. Errors are largely a function of the number of times the unit of measurement is consistently repeated within the responses comprising a person's or item's data. More items and rating scale categories result in lower errors of measurement, and more respondents result in lower calibration errors, in mathematically predictable patterns (Linacre, 1993; Woodcock, 1992).

The consistency established within the frame of reference shown in

Table 1 also makes diagnosis of unintended events in the measurement process possible. What if, for instance, one item's or one person's responses varied between correct (1) and incorrect (0) randomly, with correct and incorrect answers appearing equally often at both the easy and the difficult ends of the scale? What if a person was succeeding on the most difficult items and failing on the easy ones? What if the people with the lowest scores were succeeding on a difficult item, and the people with the highest scores were failing on it?

When data do not at least roughly conform to the pattern shown in Table 1, ambiguous questions, respondent misunderstandings, data entry errors, or other factors are introducing inconsistencies into the data that prevent the scores from being sufficient statistics. Such failures to realize the intention to measure are not reasons for abandoning the measurement effort, and the associated failure to fit a Rasch model with such data is not a justification for making do with a less exacting approach. Far from being a failure of a Rasch model to work as it should, the detection of inconsistent and insufficient response sets is the first step in what usually turns out to be successful clarification of what was previously confused.

Calculating logits

Table 1 also shows how raw scores are transformed from their original nonlinear and scale-dependent state into linear and scale-free logit measures. The logit is the log-odds unit, calculated by taking the natural logarithm of the response odds. The natural logarithm is known as a two-stretch transformation because, when applied to data distributed along a quantitative continuum, it pulls cases in the tails of the distribution further away from the center than it pulls cases in the middle of the distribution. The natural logarithm has e (2.718) as its base, and is often used in accounting, demographic studies of population growth, and economics; the bel and the bar are examples of logarithmic measurement scales. The term "two-stretch" refers to the way the log stretches out the tails of a distribution to better display the proportionate value of changes at these extremes. Economic studies of income, for instance, frequently find the distribution of their numbers clustered near zero and trailing off gradually toward very large numbers; the log is used to adjust this skew.

The value of the natural log's two stretch transformation for medical research is easy to appreciate via an example. Imagine two different experimental treatments that address two different clinical problems. The first treat-

ment would reduce the incidence of the problem by one-half of one percent, from 50% of the population to 49.5%. The second treatment also reduces the incidence of its associated problem by .5%, but in this case only 1% of the population is affected to begin with, so the reduction from 1 to .5% amounts to a 50% reduction in the incidence of the problem, whereas the first treatment effects a reduction of only 1%. The second one-half of one percent is clearly of much greater value than the first, and the logarithmic transformation makes that added value evident by the way it proportionately enlarges unit size as the extremes of the distribution are approached.

Table 2 shows how the difference between logits varies across constant differences between proportions. A difference of .0075 in the proportions is equivalent to a logit difference of 1.4 at the extremes, but is only .03 logits in the middle of the scale. Scale-free measurement incorporates these relative differences in the process of instrument calibration, and does not allow them to be forgotten.

Table 1

Hypothetical data displaying the conjoint order needed for parameter convergence and fit to a Rasch Model

	Items Easy or agreeable to hard or disagreeable												
Pers	1	2	3	4	5	6	7	8	9	10	Scor	P	Logit Meas
1	1	1	1	1	1	1	1	1	1	0	9	.9	2.20
2	1	1	1	1	1	1	1	1	0	1	9	.9	2.20
3	1	1	1	1	1	1	1	0	1	0	8	.8	1.39
4	1	1	1	1	1	1	0	0	1	0	7	.7	0.85
5	1	1	1	1	0	1	1	1	0	0	7	.7	0.85
6	1	1	1	1	0	1	0	0	0	0	5	.5	0.00
7	1	1	1	0	1	0	0	0	0	0	4	.4	-0.41
8	1	1	0	0	1	0	0	0	0	0	3	.3	-0.85
9	0	1	1	1	0	0	0	0	0	0	3	.3	-0.85
10	1	0	1	0	0	0	0	0	0	0	2	.2	-1.39
Item													
Score	9	9	9	7	6	6	4	3	3	1			
Scor2	1	1	1	3	4	4	6	7	7	9			
P	.1	.1	.1	.3	.4	.4	.6	.7	.7	.9			
Logit Calib	-2.2	-2.2	-2.2	-0.9	-0.4	-0.4	0.4	0.9	0.9	2.2			

Scores are sums of the 1s and 0s. In order to estimate the difficulty of the items, the score is converted into a count of the number of persons failing (Scor2). Proportions (P) are the scores divided by the number of items for the persons, and by the number of persons for the items (both are 10 in this example). The logit calibrations (Logit Calib) and the logit measures (Logit Meas) are the natural logarithm of the odds (P/1-P) of failure for the items, and of success for the persons. For more information, see Chapter 2 in Wright and Stone (1979, pp. 28-45), or Bond and Fox (2001)..

The calculations shown in Table 1 display how missing data is accounted for in objective measurement. Because the number of items administered is included in the denominator when calculating the proportions P, it can vary across persons with no effect on the linearity of the unit of measurement. The number of items administered is also included in the error calculation, and measurement error rises as the number of items decreases.

Equating for universal metrics

For two or more instruments intended to measure the same construct, an experimental test of the potential for equating can be undertaken if it is possible to code the data from each instrument to point in the same direction, so that a higher rating means the same thing (typically, more functional independence, satisfaction, etc.) for the data from each instrument. Partial credit models (Wright and Masters, 1982; Masters, 1982; McArthur, Casey, Morrow, et al., 1992; Zhu and Kurz, 1994) implemented by many of the available computer programs (see above) make it possible for the instruments to employ different numbers of rating categories. The number of categories might also vary across items within an instrument, and even if the instruments or items do have the same number of categories, they do not have to mean exactly the same thing, as long as they are consistently coded so that a higher rating means more independence, for instance.

With this observational framework in place, then the instruments must be used to produce data from a common sample of patients. If there are more

Table 2

Logits from Proportions

Proportion	Logit		Proportion	Logit
.0025	-5.989		.9975	5.989
.01	-4.595		.99	4.595
.02	-3.892		.98	3.892
.05	-2.944		.95	2.944
.10	-2.197		.90	2.197
.25	-1.099		.75	1.099
.49	-.040		.51	.040
.4925	-.030		.5075	.030
.50	.000		.50	.000

than two instruments, it is not necessary to administer all of them to every patient, but each instrument must be linked to every other, if not through a common sample then through common equating to a third set of items.

The data from each instrument must then fit a probabilistic measurement model and generate stable item scale values that can be set (anchored) at their positions on the common measurement continuum. When data from items that were not designed to fit a probabilistic model are analyzed in this way, substantial problems in instrument quality often need to be addressed before it is possible to move on to the creation of scale-free measurement standards. Using scatter plots to study the relative performances of the instruments on the common sample rated illuminates their respective qualities, which aids in evaluating their measurement characteristics and the quantitative properties of the variable (Bland and Altman, 1986; Ottenbacher and Stull, 1993; Wright and Masters, 1982).

When data from several instruments fit a common measurement model, and so are shown to measure the same thing, then it is reasonable to treat all of the items as coming from a single bank and to equate them on a common metric. When the quality of this calibration is satisfactory, then the individual instruments' item scale values can be anchored at their cocalibration positions to produce measures in the metric of the scale-free measurement standard (Cella, et al., 1996; Fisher, 1997a, 1997b; Fisher, Harvey, Taylor, et al., 1995; Fisher, Eubanks, and Marier, 1997; Fisher, Harvey, and Kilgore, 1995; Gonin, et al., 1996; Segal, Heinemann, Schall, et al., 1997; Tennant and Young, 1997).

Moving into the future

These procedures are not difficult to implement, but are not yet widely employed in classrooms, clinics, and research. If they were better known, McDowell and Newell (1996, p. 80) would not have to lament the fact that

> the development of ADL scales has been so uncoordinated. Many scales have not been planned on a systematic review of the strengths and weaknesses of previous instruments, and the definition of disability itself is more often assumed than clearly stated. The application of these instruments does not seem to have led to a cumulative understanding of the concept of disability, its relation to impairment and handicap, or of the sequence in which changes in disability occur as a patient's condition changes.... We still know relatively little about the overlap among the various measurement methods, and the few comparative studies that exist mainly review the older scales.

These flaws can be corrected by exploring the extent to which various functional status measures involve the same constructs, and so can be cocalibrated, or equated. Cocalibration is a coordination of existing instruments that will elucidate their relative strengths and weaknesses; more clearly define disability; accumulate a body of scientific knowledge on disability; facilitate the study of disability in relation to impairment and handicap; and improve communication about disability by providing a universal metric that can function as a common currency for the exchange of quantitative information.

It is often said that if you cannot measure, you cannot manage. In a world where every brand of functional status instrument has its own idiosyncratic measuring unit, and few persons employing these instruments have a grasp of, or are able to apply, scale-free amounts of functional status in their work, there is probably little actual measurement or management taking place. Successful applications of probabilistic measurement models to functional assessment data introduce reasons for believing that populations of persons with disabilities can be matched with relevant scale-free item hierarchies.

Research is progressing by identifying common hierarchies of similar items from different instruments calibrated on different samples (Fisher, 1997a), as well as by evaluating data from two or more similar instruments administered to a common sample (Cella, et al., 1996; Fisher, Harvey, and Kilgore, 1995; Fisher, Harvey, Taylor, et al., 1995; Fisher, Eubanks, and Marier, 1997; Gonin, et al., 1996; Grimby, et al., 1996; Segal, Heinemann, Schall, et al., 1997; Tennant and Young, 1997; Zhu, 1996). Also in progress is the drafting of a *Proposed Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Psychosocial Test Method* (Fisher, 1997c), and an accompanying glossary of measurement terms (Fisher, 1996)[4], both modeled after the ASTM E - 691 -92 standard practice and its associated documents.

The vast majority of functional assessment data Rasch analyses to date have been conducted in isolation from information on similar analyses of similar instruments. This trend could continue until uniformities among the item hierarchies is so obvious that not equating the instruments would be foolish. On the other hand, if researchers would incorporate comparisons of the results of their Rasch analyses with the results of prior work, progress toward the goals spelled out by McDowell and Newell (1989)

would accelerate. After all, can any claim to objective measurement that does not result in systems for applying nonarbitrary scale-free metrics at any given relevant point of use really be considered valid and complete?

Footnotes

[1] Legend holds that in developing the method of paired comparisons and coming so close to scale-free measurement, Thurstone stole fire from the gods. In retribution they chained him to factor analysis (Lumsden, 1976).

[2] Common item sets and sample sizes are not necessary for successful fit to a Rasch model, but are assumed in this statement for the sake of simplicity.

[3] Both are available from the author on request.

Acknowledgements

This is an extensively revised version of a paper presented to the Conference on Rehabilitation Outcome Analysis: State of the Art Outcomes by Level of Care, presented by the Rehabilitation Foundation, Inc., in cooperation with Marianjoy Rehabilitation Hospital and Clinics, November 15-16, 1994, in Oak Brook, Illinois. I would like to thank Richard F. Harvey, Robert L. Marier, James H. Diaz, Miguel Guzman, and the late Colonel Gordon Black for their support of and influence on this work. Theo Dawson, Norbert Goldfield, and John Wasson provided extensive comments on earlier versions of this paper; I'd like to thank them for their input, even if I was unable to always incorporate their perspectives. None of what follows would have been possible without Benjamin Wright's inspiring illuminations, but any and all flaws are my own responsibility.

References

Ackermann, J. R. (1985). *Data, instruments, and theory: A dialectical approach to understanding science.* Princeton, New Jersey: Princeton University Press.

Adams, R. J., and Kboo, S. (1995). *Quest: The interactive Rasch test analysis system, version 2.* Camberwell, Victoria, Australia: Australian Council for Educational Research. (Distributed in North America by Assessment Systems Corp., St. Paul, MN).

Adams, R. J., and Wright, B. D. (1994). When does misfit make a difference? In M. Wilson (Ed.), *Objective Measurement: Theory into Practice, Volume 2* (pp. 244-270). Norwood, NJ: Ablex.

Allerup, P., and Sorber, G. (1977). *The Rasch model for questionnaires. Report no. 16.* Copenhagen, Denmark: Danish Institute for Educational Research.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*(1), 69-81.

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

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika, 43,* 357-374.

Andrich, D. (1978b). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2,* 449-460.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. B. Tuma (Ed.), *Sociological methodology 1985* (pp. 33-80). San Francisco: Jossey-Bass.

Andrich, D. (1988). *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.

Andrich, D. (1990, Summer). The ability of an item. *Rasch Measurement Transactions, 4*(2), 101-102.

Andrich, D. (1997). *RUMM: Rasch unidimensional models for measurement.* Software for Windows. School of Education, Murdoch University, Australia.

Andrich, D., Lynne, A., and Sheridan, B. (1990). *ASCORE: A Fortran IV program for analyzing psychometric responses according to a Rasch binomial logistic model.* Murdoch, Western Australia: School of Education, Murdoch University.

ASC. (1996). *RASCAL: Rasch analysis program.* St. Paul, MN: Assessment Systems Corp.

ASTM Committee E-11 on Statistical Methods. (1992). *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.

ASTM Committee E-31 on Computerized Systems. (1996). *Standard Guide for Content and Structure of the Computer-Based Patient Record (E 1384 - 96).* West Conshohocken, PA: American Society for Testing and Materials.

Bernstein, R. J. (1983). *Beyond objectivism and relativism: Science, hermeneutics, and praxis.* Philadelphia: University of Pennsylvania Press.

Bland, J. M., and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 1,* 307-310.

Board of Directors of the American Medical Informatics Association. (1994). Standards for medical identifiers, codes, and messages needed to create an efficient computer-stored medical record. *Journal of the American Medical Informatics Association, 1*(1), 1-7.

Bond, T., and Fox, C. (2001). *Fundamental measurement in the human sciences: Applying the Rasch model.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Borman, W. C. (1978). Exploring the upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology, 63,* 134-144.

Bradley, R. A., and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of pair comparisons. *Biometrika, 63,* 324-345.

Brink, N. E. (1972). Rasch's logistic model vs the Guttman model. *Educational and Psychological Measurement, 32,* 921-927.

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika, 42,* 631-634.

Brown, H. I. (1977). *Perception, theory and commitment: The new philosophy of science.* Chicago, Illinois: University of Chicago Press.

Bud, R., and Cozzens, S. E. (Editors). (1992). *SPIE Institutes. Vol. 9: Invisible connections: instruments, institutions, and science* (R. F. Potter, Ed.). Bellingham, WA: SPIE Optical Engineering Press.

Campbell, N. R. (1920). *Physics, the Elements.* Cambridge: Cambridge University Press.

Cason, G. J., and Cason, C. L. (1984). A deterministic theory of clinical performance rating. *Evaluation and the Health Professions, 7,* 221-247.

Cella, D. F., Lloyd, S. R., and Wright, B. D. (1996). Cross-cultural instrument equating: Current research and future directions. In B. Spilker (Ed.), *Quality of life and pharmacoeconomics in clinical trials (2d edition)* (pp. 707-715). New York, New York: Lippincott-Raven.

Choppin, B. (1976). Recent developments in item banking. In D. N. DeGruitjer, and L. J. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement* (pp. 233-245). New York: Wiley.

Cliff, N. (1973). Scaling. *Annual Review of Psychology, 24,* 473-526.

Costich, J. (1993, August). Assessment of measurement practices in medical rehabilitation. *Physical Medicine and Rehabilitation Clinics of North AmericaPhysical Medicine and Rehabilitation Clinics of North America: New Developments in Functional Assessment, 4*(3), 587-594.

Daston, L. (1992). Baconian facts, academic civility, and the prehistory of objectivity. *Annals of Scholarship,* 8, 337-363.

Daston, L., and Galison, P. (1992, Fall). The image of objectivity. *Representations, 40,* 81-128.

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical.* New York: Russell Sage Foundation.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistics Society, 53,* 460-475 and 644-663.

Englehard, G., Jr. (1984). Thorndike, Thurstone, and Rasch: A comparison of their methods of scaling psychological and educational tests. *Applied Psychological Measurement, 8*(1), 21-38.

Englehard, G., Jr. (1992). The measurement of writing ability with a many-facet Rasch model. *Applied Measurement in Education, 5*(3), 171-191.

Englehard, G., Jr. (1994). Historical views of the concept of invariance in measurement theory. In M. Wilson (Ed.), *Objective Measurement: Theory into Practice, Vol. 2* (pp. 73-99). Norwood, NJ: Ablex.

Esdaille, M., Shaw, F., Smith, M., and Valgeirsdóttir, S. (1994). Educational applications of probabilistic conjoint measurement models. *International Journal of Educational Research, 21*(6), 635-651.

Fischer, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Fischer, G. (1974). *Einfuhrungin die theorie psychologischer tests.* Vienna: Verlag Hans Huber.

Fischer, G., and Molenaar, I. (1995). *Rasch models: Foundations, recent developments, and applications.* New York: Springer-Verlag.

Fisher, A. G. (1993, April). The assessment of IADL motor skills: An application of many-faceted Rasch analysis. *American Journal of Occupational Therapy, 47*(4), 319-329.

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

Fisher, A. G., Bryze, K. A., Granger, C. V., Haley, S. M., Hamilton, B. B., Heinemann, A. W., Puderbaugh, J. K., Linacre, J. M., Ludlow, L. H., McCabe, M. A., and Wright, B. D. (1994a). Applications of conjoint measurement to the development of functional assessments. *International Journal of Educational Research, 21*(6), 579-593.

Fisher, P. B. (1993). Measuring ability in sports performance: The case of baseball [MA thesis]. Chicago, Illinois: University of Chicago.

Fisher, P. B. (1995, April). *Measuring baseball performance.* Unpublished ms. International Objective Measurement Workshops, University of California - Berkeley.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A, 222,* 309-368.

Fisher, W. P., Jr. (1988). Truth, method, and measurement: The hermeneutic of instrumentation and the Rasch model [Diss]. *Dissertation Abstracts Inter-*

national, 49, 0778A, Chicago, Illinois: University of Chicago (376 pages, 23 figures, 31 tables).

Fisher, W. P., Jr. (1991, April). *The hermeneutic of additive conjoint measurement in educational research.* [ERIC Document #TM 016 361]. American Educational Research Association. Chicago.

Fisher, W. P., Jr. (1992). Objectivity in measurement: A philosophical history of Rasch's separability theorem. In M. Wilson (Ed.), *Objective measurement: Theory into practice. Vol. I* (pp. 29-58). Norwood, New Jersey: Ablex Publishing Corporation.

Fisher, W. P., Jr. (1993). Measurement-related problems in functional assessment. *The American Journal of Occupational Therapy, 47*(4), 331-338.

Fisher, W. P., Jr. (1994). The Rasch debate: Validity and revolution in educational measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice. Vol. II* (pp. 36-72). Norwood, New Jersey: Ablex Publishing Corporation.

Fisher, W. P., Jr. (1996, October). *Rating scale measurement standards relevant to ASTM 1384 on the content and structure of the electronic health record.* Unpublished paper. ASTM E31 Committee on the Electronic Health Record, Washington, DC.

Fisher, W. P., Jr. (1997a). Physical disability construct convergence across instruments: Towards a universal metric. *Journal of Outcome Measurement, 1*(2), 87-113.

Fisher, W. P., Jr. (1997b, October). What scale-free measurement means to health outcomes research. *Physical Medicine and Rehabilitation State of the Art Reviews, 11*(2), 357-373.

Fisher, W. P., Jr. (1997c, April). *Draft rating scale measurement standard procedure.* Unpublished paper. ASTM E31 Committee on the Electronic Health Record, Nashville.

Fisher, W. P., Jr., Eubanks, R. L., and Marier, R. L. (1997). Equating the MOS SF36 and the LSU HSI physical functioning scales. *Journal of Outcome Measurement, 1*(4), 329-362.

Fisher, W. P., Jr., and Fisher, A. G. (1993). Applications of Rasch analysis to studies in occupational therapy. *Physical Medicine and Rehabilitation Clinics of North America, 4*(3), 551-569.

Fisher, W. P., Jr., Harvey, R. F., and Kilgore, K. M. (1995). New developments in functional assessment: Probabilistic models for gold standards. *NeuroRehabilitation, 5*(1), 3-25.

Fisher, W. P., Jr., Harvey, R. F., Taylor, P., Kilgore, K. M., and Kelly, C. K. (1995). Rehabits: A common language of functional assessment. *Archives of Physical Medicine and Rehabilitation, 76,* 113-122.

Fisher, W. P., Jr., Vial, R. H., and Sanders, C. V. (1997, May). Removing rater effects from a medical clerkship evaluation: A Facets analysis. *Academic Medicine, 72*(5), 443-44.

Fisher, W. P., Jr., and Wright, B. D. (1994). Introduction to probabilistic conjoint measurement theory and applications. *International Journal of Educational Research, 21*(6), 559-568.

Glas, C. A. W., and Ellis, J. L. (1995). *RSP 1.0: Rasch scaling program.* Arnhem, Netherlands: CITO.

Gonin, R., Lloyd, S. R., and Cella, D. F. (1996). Establishing equivalence between scaled measures of quality of life. *Quality of Life Research, 5*, 20-26.

Green, K. (1986). Fundamental measurement: A review and application of additive conjoint measurement in educational testing. *Journal of Experimental Education, 54*(3), 141-147.

Grimby, G., Andrén, E., Holmgren, E., Wright, B., Linacre, J. M., and Sundh, V. (1996, November). 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, 77*(11), 1109-1114.

Guilford, J. P. (1936). *Psychometric methods.* New York: McGraw-Hill.

Gulliksen, H. (1946). Paired comparisons and the logic of measurement. *Psychological Review, 53,* 199-213.

Gustafsson, J. (1979). *PML: A computer program for conditional estimation and testing in the Rasch model for dichotomous items. Report no. 85.* Göteborg, Sweden: Institute of Education, University of Göteborg.

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

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer and et al. (Eds.), *Studies in social psychology in World War II. volume 4: Measurement and prediction* (pp. 60-90). New York: Wiley.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science.* Cambridge: Cambridge University Press.

Hamilton, B. B., Granger, C. V., Sherwin, F. S., Zielezny, M., and Tashman, J. (1987). A uniform national data system for medical rehabilitation. In M. J. Fuhrer (Ed.), *Rehabilitation Outcomes* (pp. 137-). Baltimore: Paul H. Brookes.

Harvey, R. F., and Jellinek, H. M. (1981). Functional performance assessment: A program approach. *Archives of Physical Medicine and Rehabilitation, 62,* 456-461.

Heelan, P. (1965). *Quantum mechanics and objectivity: A study in the physical philosophy of Werner Heisenberg.* The Hague: Martinus Nijhoff.

Heelan, P. (1972). Towards a hermeneutic of natural science. *Journal of the British Society for Phenomenology, 3,* 252-260.

Heelan, P. (1983a, June). Natural science as a hermeneutic of instrumentation. *Philosophy of Science, 50,* 181-204.

Heelan, P. (1983b, September). Perception as a hermeneutical act. *Review of Metaphysics, 37,* 61-75.

Heelan, P. (1983c). *Space perception and the philosophy of science.* Berkeley, California: University of California Press.

Heelan, P. (1993, October). Theory of social-historical phenomena: Quantum mechanics and the social sciences. Society for the Philosophy of the Human Sciences. New Orleans, LA.

Heidegger, M. (1967). *What is a thing?* (W. B. Barton, Jr. and V. Deutsch, Trans.). South Bend, Indiana: Regnery/Gateway.

Heinemann, A. W., Hamilton, B. B., Granger, C. V., Wright, B. D., Linacre, J. M., Betts, H. B., Aguda, B., and Mamott, B. D. (1991). *Rating scale analysis of functional assessment measures* [NIDRR Innovation Grant Award Final Report]. Chicago, Illinois: Rehabilitation Institute of Chicago.

Heinemann, A. W., Linacre, J. M., Wright, B. D., Hamilton, B. B., and Granger, C. V. (1993). Relationships between impairment and physical disability as measured by the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation, 74*(6), 566-573.

Heinemann, A. W., Linacre, J. M., Wright, B. D., Hamilton, B. B., and Granger, C. V. (1994). Prediction of rehabilitation outcomes with disability measures. *Archives of Physical Medicine and Rehabilitation, 75*(2), 133-143.

Hess, R. K. (1995, April). *Measuring school test scores.* Unpublished ms. International Objective Measurement Workshop, Berkeley, California.

Hesse, M. (1972). In defence of objectivity. *Proceedings of the British Academy, 58,* 275-292.

Holm, K., and Kavanagh, J. (1985). An approach to modifying self-report instruments. *Research in Nursing and Health, 8,* 13-18.

Ihde, D. (1991). *Instrumental realism: The interface between philosophy of science and philosophy of technology.* The Indiana Series in the Philosophy of Technology. Bloomington, Indiana: Indiana University Press.

Keats, J. A. (1967). Test theory. *Annual Review of Psychology, 18,* 217-238.

Kelderman, H. (1986). *Common item equating using the log-linear Rasch model.* Twente, Netherlands: Department of Education, University of Twente.

Kelderman, H., and Steen, R. (1988). *LOGIMO computer program for log-linear item response theory modelling.* Twente, Netherlands: Department of Education, University of Twente.

Kilgore, K. M., Fisher, W. P., Jr., Silverstein, B., Harley, J. P., and Harvey, R. F. (1993). Application of Rasch analysis to the Patient Evaluation and Conference System. *Physical Medicine and Rehabilitation Clinics of North America, 4*(3), 493-515.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971). *Foundations of measurement. Volume 1: Additive and polynomial representations.* New York: Academic Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

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

Latour, B. (1989). *Science in Action.* Cambridge, MA: Harvard University Press.

Linacre, J. M. (1989). *Many-facet Rasch measurement.* Chicago: MESA Press.

Linacre, J. M. (1991a, Summer). Beyond partial credit. *Rasch Measurement Transactions, 5*(2), 155.

Linacre, J. M. (1991b, April). Structured rating scales [ERIC TM 016615]. International Objective Measurement Workshops. Chicago.

Linacre, J. M. (1993). Rasch generalizability theory. *Rasch Measurement Transactions, 7*(1), 283-284.

Linacre, J. M. (1995). *Facets Rasch analysis computer program.* Chicago: MESA Press.

Linacre, J. M. (1998). *WINSTEPS Rasch analysis computer program.* Chicago: MESA Press.

Linacre, J. M., Englehard, G., Tatum, D. S., and Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research, 21*(6), 569-577.

Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V., and Hamilton, B. B. (1994). The structure and stability of the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation, 75*(2), 127-132.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs, 61*(4:#285).

Lord, F. M. (1983). Small *N* justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 51-61). New York, NY: Academic Press, Inc.

Luce, R. D., and Tukey, J. W. (1964). Simultaneous conjoint measurement: A

new kind of fundamental measurement. *Journal of Mathematical Psychology, 1*(1), 1-27.

Ludlow, L. H. (1985). A strategy for the graphical representation of Rasch model residuals. *Educational and Psychological Measurement, 45,* 851-859.

Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27,* 1-280.

Lumsden, J. (1980). Variations on a theme by Thurstone. *Applied Psychological Measurement, 4*(1), 1-7.

Lunz, M. E., Bergstrom, B. A., and Gershon, R. C. (1994). Computer adaptive testing. *International Journal of Educational Research, 21*(6), 623-634.

Lunz, M. E., and Stahl, J. A. (1993a, April). The effect of rater severity on person ability measures: A Rasch model analysis. *American Journal of Occupational Therapy, 47*(4), 311-317.

Lunz, M. E., and Stahl, J. A. (1993b). Impact of examiners on candidate scores: An introduction to the use of multifacet Rasch model analysis for oral examinations. *Teaching and Learning in Medicine, 5*(3), 174-181.

Lunz, M. E., Stahl, J. A., and Wright, B. D. (1996). The invariance of judge severity calibrations. In G. Englehard and M. Wilson (Eds.), *Objective Measurement: Theory into Practice, Volume 3* (pp. 99-112). Norwood, NJ: Ablex.

Lunz, M. E., Wright, B. D., and Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3/4,* 331-345.

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

Mandel, J. (1978, December). Interlaboratory testing. *ASTM Standardization News, 6,* 11-12.

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

Masters, G. N. (1985, March). Common-person equating with the Rasch model. *Applied Psychological Measurement, 9*(1), 73-82.

McArthur, D. L., Casey, K., Morrow, T. J., Cohen, M. J., and Schandler, S. L. (1992). Partial-credit modeling and response surface modeling of biobehavioral data. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (pp. 109-120). Norwood, New Jersey: Ablex.

McDowell, I., and Newell, C. (1989). *Measuring health: A guide to rating scales and questionnaires.* Oxford: Oxford University Press.

McDowell, I., and Newell, C. (1996). *Measuring health: A guide to rating scales and questionnaires. 2d edition.* Oxford: Oxford University Press.

Mendelsohn, E. (1992). The social locus of scientific instruments. In R. Bud and S. E. Cozzens (Eds.), *Invisible connections: Instruments, institutions, and science* (pp. 5-22). Bellingham, WA: SPIE Optical Engineering Press.

Merbitz, C., Morris, J., and Grip, J. (1989). Ordinal scales and the foundations of misinference. *Archives of Physical Medicine and Rehabilitation, 70,* 308-312.

Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*(3), 355-383.

Myford, C. M. (1989). The nature of expertise in aesthetic judgment: Beyond inter-judge agreement [Diss]. *Dissertation Abstracts International, 50,* 3562A, Chicago, Illinois: University of Chicago.

Myford, C. M., Marr, D. B., and Linacre, J. M. (1996). *Reader calibration and its potential role in equating for the Test of Written English.* Center for Performance Assessment, vol. 95-02. Princeton, NJ: Educational Testing Service (69 pages).

Myford, C. M., and Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system.* Center for Performance Assessment, vol. 94-05. Princeton, NJ: Educational Testing Service (88 pages).

Nicholson, G. (1984). *Seeing as reading.* Atlantic Highlands, NJ: Humanities Press.

O'Connell, J. (1993). Metrology: The creation of universality by the circulation of particulars. *Social Studies of Science, 23,* 129-173.

Ottenbacher, K., and Stull, G. (1993). The analysis and interpretation of method comparison studies in rehabilitation research. *American Journal of Physical Medicine and Rehabilitation, 72,* 266-271.

Perline, R., Wright, B. D., and Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*(2), 237-255.

Plato. (1961). *Theaetetus* (F. M. Cornford, Trans.). In E. Hamilton and H. Cairns (Eds.), The Collected Dialogues of Plato, including the Letters. Bollingen Series LXXI. Princeton, NJ: Princeton University Press.

Price, D. J. de Solla. (1986). Of sealing wax and string. In D. J. de Solla Price, *Little Science, Big Science—and Beyond.* New York: Columbia University Press.

Rasch, G. (1960). *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.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (pp. 321-333). Berkeley, California: University of California Press.

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

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8,* 3.

Rehfeldt, T. K. (1990, Autumn). Directness and measurement. *Rasch Measurement Transactions, 4*(3), 117.

Ricoeur, P. (1977). *The rule of metaphor: Multi-disciplinary studies of the creation of meaning in language* (R. Czerny, Trans.). Toronto: University of Toronto Press.

Ruch, G. M. (1929). *The objective or new-type examination.* Chicago: Scott, Foresman.

Ruggles, A. M. (1911). *Grades and grading.* New York: Teacher's College Press.

Schaffer, S. (1992). Late Victorian metrology and its instrumentation: A manufactory of Ohms. In R. Bud and S. E. Cozzens (Eds.), *Invisible connections: Instruments, institutions, and science* (pp. 23-56). Bellingham, WA: SPIE Optical Engineering Press.

Segal, M., Heinemann, A., Schall, R. R., and Wright, B. D. (1997, June). Rasch analysis of a brief physical ability scale for long-term outcomes of stroke. *Physical Medicine and Rehabilitation State of the Art Reviews, 11*(2), 385-396.

Shapin, S. (1989, November-December). The invisible technician. *American Scientist, 77,* 554-563.

Silverstein, B. J., Kilgore, K. M., and Fisher, W. P., Jr. (1989). Implementing patient tracking systems and using functional assessment scales. Center for Rehabilitation Outcome Analysis monograph series on issues and methods in rehabilitation outcome analysis, Volume 1. Wheaton, IL: Marianjoy Rehabilitation Center.

Silverstein, B. J., Fisher, W. P., Jr., Kilgore, K. M., Harvey, R. F., and Harley, J. P. (1992). Applying psychometric criteria to functional assessment in medical rehabilitation: II. defining interval measures. *Archives of Physical Medicine and Rehabilitation, 73*(6), 507-518.

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

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

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

Smith, R. M. (1994). A comparison of the power of Rasch total and between fit statistics to detect measurement disturbances. *Educational and Psychological Measurement, 54,* 42-55.

Smith, R. M. (1998, May 15-16). PECS/FIM equating. Second International Outcome Measurement Conference, University of Chicago.

Smith, R. M., and Kramer, G. A. (1992). A comparison of two methods of test equating in the Rasch model. *Educational and Psychological Measurement, 52,* 835-846.

Smith, R., Julian, E., Lunz, M., Stahl, J., Schulz, M., and Wright, B. D. (1994). Applications of conjoint measurement in admission and professional certification programs. *International Journal of Educational Research, 21*(6), 653-664.

Stahl, J. A., and Lunz, M. E. (1996). Judge performance reports: Media and message. In G. Englehard and M. Wilson (Eds.), *Objective Measurement: Theory into Practice, Volume 3* (pp. 113-125). Norwood, NJ: Ablex.

Stent, G. (1981). Cerebral hermeneutics. *Journal of Social and Biological Structures, 4,* 107-124.

Streiner, D. L., and Norman, G. R. (1995). *Health Measurement Scales: A Practical Guide to their Development and Use, 2d edition.* New York: Oxford University Press.

Stucki, G., Daltroy, L., Katz, N., Johannesson, M., and Liang, M. H. (1996). 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, 49*(7), 711-717.

Tatum, D. S. (1991). A measurement system for speech evaluation [Diss]. *Dissertation Abstracts International, 52*(4), 1301A, Chicago, Illinois: University of Chicago.

Tennant, A., and Young, C. (1997, June). Coma to community: Continuity in measurement. *Physical Medicine and Rehabilitation State of the Art Reviews, 11*(2), 375-384.

Thurstone, L. L. (1959). *The measurement of values.* Chicago: University of Chicago Press, Midway Reprint Series.

Todes, S. J., and Dreyfus, H. L. (1970). The existentialist critique of objectivity. In J. M. Edie, F. H. Parker and C. O. Scrag (Eds.), *Patterns of the life-world: Essays in honor of John Wild* (pp. 346-387). Evanston, IL: Northwestern University Press.

van der Linden, W. (1994). Fundamental measurement and the fundamentals of Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (pp. 3-24). Norwood, NJ: Ablex Publishing Corporation.

Verhelst, N. D. (1993). *One Parameter Logistic Model computer program.* Arnhem, Netherlands: Cito.

Verhelst, N. D., and Glas, C. A. W. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58*(3), 395-415.

Wegener, B. (1982). Introduction: Outline of a structural taxonomy of sensory and social psychophysics. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 1-42). Hillsdale, NJ: Lawrence Erlbaum.

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

Weiss, D. J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Weiss, D. J., and Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361-375.

Wernimont, G. (1977, March). Ruggedness evaluation of test procedures. *ASTM Standardization News, 5,* 13-16.

Wernimont, G. (1978, December). Careful intralaboratory study must come first. *ASTM Standardization News, 6,* 11-12.

Widmalm, S. (1995, May 12). On exactitude, a review of *The Values of Precision,* by M. Norton Wise. *Science, 268,* 905-6.

Wilson, M. (1989). A comparison of deterministic and probabilistic approaches to learning structures. *Australian Journal of Education, 33*(2), 127-140.

Wilson, T. P. (1971). Critique of ordinal variables. *Social Forces, 49,* 432-444.

Wise, M. N. (1988). Mediating machines. *Science in Context, 2*(1), 77-113.

Woodcock, R. W. (1992). Woodcock test design nomograph. *Rasch Measurement Transactions, 6*(3), 243-244.

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

Wright, B. D. (1977a). Misunderstanding the Rasch model. *Journal of Educational Measurement, 14*(3), 219-225.

Wright, B. D. (1977b). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97-116.

Wright, B. D. (1980). Foreword, Afterword. In *Probabilistic models for some intelligence and attainment tests, by Georg Rasch* [Reprint; original work published in 1960 by the Danish Institute for Educational Research]. Chicago: University of Chicago Press.

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review, 3*(1), 281-288.

Wright, B. D. (1985). Additivity in psychological measurement. In E. Roskam (Ed.), *Measurement and personality assessment.* North Holland: Elsevier Science Ltd.

Wright, B. D. (1993, Summer). Equitable test equating. *Rasch Measurement Transactions, 7*(2), 298-299.

Wright, B. D. (1997, Winter). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33-45, 52.

Wright, B. D., and Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement, 21*(4), 331-345.

Wright, B. D., and Douglas, G. A. (1975). *Best test design and self-tailored testing.* Research Memorandum, vol. 19. MESA Laboratory, Department of Education, University of Chicago.

Wright, B. D., and Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*(12), 857-867.

Wright, B. D., Linacre, J. M., and Heinemann, A. W. (1993). Measuring functional status in rehabilitation. In C. V. Granger and G. E. Gresham (Eds.), *New developments in functional assessment. Physical Medicine and Rehabilitation Clinics of North America, 4*(3), 475-491

Wright, B. D., and Linacre, J. M. (1995). *A User's Guide to BIGSTEPS Rasch-Model Computer Program.* Chicago: MESA Press.

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

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

Wu, M. L., Adams, R. J., and Wilson, M. R. (1996). *MATS: Multi-aspect test software.* Camberwell, Victoria, Australia: Australian Council for Educational Research.

Zhu, W. (1996). Should total scores from a rating scale be used directly? *Research Quarterly for Exercise and Sport, 67*(3), 363-372.

Zhu, W., and Kurz, K. A. (1994). Rasch partial credit analysis of gross motor competence. *Perceptual and Motor Skills, 79,* 947-961.

Zupko, R. E. (1977). *British weights and measures: A history from antiquity to the seventeenth century.* Madison, WI: University of Wisconsin Press.