DOCUMENT RESUME

ED 331 857 TM 016 378

AUTHOR Fisher, William P., Jr.
TITLE The Rasch Debate: Validity and Revolution in Educational Measurement.
PUB DATE Apr 91
NOTE 44p.; Paper presented at the International Objective Measurement Workshop (6th, Chicago, IL, April 1991).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Construct Validity; Content Validity; *Educational Assessment; *Item Response Theory; Literature Reviews; *Measurement Techniques; *Test Construction; Test Items; Test Validity
IDENTIFIERS *Rasch Model

ABSTRACT

In an address to the National Council on Measurement in Education, R. M. Jaeger (1987) commented that there appears to be a fundamental difference in measurement philosophy between those on the two sides of the debate over the Rasch model. Jaeger's observations are explicated by contrasting the views on measurement of B. D. Wright and E. F. Lindquist with relation to the interpretation each has of the validation of constructs as considered by C. Cherryholmes (1988). Lindquist conceives of test items as given in an objective reality; the discursive action of construct validation is assumed to take place outside the context in which the construct is manifest. Wright suggests that test items amount to nothing more than guesses about how a construct articulates itself. Instead of objectifying test takers by subjecting them to unquestionable authority, the Rasch approach to test construction suggests that questions are tested by the respondents just as much as the respondents are tested by the questions. It is argued that by recognizing the inevitability of the imposition of political, moral, and aesthetic criteria on test items and data, and by formulating models of how these criteria can be implemented and criticized, the Rasch model and supporters of Wright's position have made a great contribution to measurement. Two tables expand on the discussion. A 133-item list of references is included. (SLD)

The Rasch Debate:
Validity and Revolution in Educational Measurement

William P. Fisher, Jr., Research Associate
Marianjoy Rehabilitation Center
Wheaton, IL 60189
708/462-4277

Paper presented to
The Sixth International Objective Measurement Workshops
University of Chicago
April, 1991

DRAFT OF WORK IN PROGRESS -- NOT FOR CITATION

CONTENTS

Objectives. In his Presidential Address to the National Council on Measurement in Education, Richard Jaeger (1987, p. 8) commented that "there appears to be a fundamental difference in measurement philosophy between those on the two sides of the Rasch debate." Jaeger then uses two quotes, one each from Benjamin Wright and E.F. Lindquist, to elaborate upon the two primary ways measurement is approached in American educational research. The purpose of this paper is to explicate Jaeger's observations by contrasting Wright and Lindquist's measurement philosophies in terms of the discourses of research (Cherryholmes, 1988) that each brings to the validation of constructs.

Theoretical framework. The validation of constructs is a discursive and investigative activity that is prior in importance to other needs for validation in social research (Cronbach & Meehl, 1955; Messick, 1975; Cherryholmes, 1988). Construct validity is discursive in focusing on the meaning of what is measured and in acknowledging that this meaning emerges from a network of interrelations. Because traditional research methodologies largely have refrained from addressing problems of meaning and interpretation, Cherryholmes explores construct validity from the perspectives of alternative research traditions, including phenomenology, ethnography, critical theory, interpretive analytics, and deconstruction.

These alternative research traditions are more explicitly focused on the fluidity of meaning, the fact that observers are always already caught up in the flow of meaning by the time the quality or the direction of that flow is recognized. Being caught up in the flow of meaning does not mean that observers are unable to alter its quality or direction, but it does mean that the validity of constructs cannot be determined by definitions, objects, or methods, which are the basic staples of traditional approaches in social research. Cronbach and Meehl (1955) start from a traditional approach, and, as Cherryholmes (1988, p. 423) points out, they are explicitly positivist. Positivism holds that science needs no prior sense of the whole to which parts belong in order to observe the parts: the roots of this position can be traced at least to Pythagoras but it reached its fullest

development in Newton's effort to eliminate hypotheses and metaphysics from science (Burtt, 1954, p. 222). The historical distance between us and Newton allows us to see that even the facts of his physics were recognized only in light of the particular ideological, political, and economic influences that served to focus his attention and that of his colleagues. These "systems of intuitive meaning" (Hudson, 1972, p. 160), tacit understandings (Polanyi, 1958, 1967), fore-structures of understanding (Heidegger, 1962), disciplinary matrices or paradigms (Kuhn, 1970), and historically-effective consciousnesses (Gadamer, 1989) focus attention selectively and create the context for the manifestation of the material form of images (Lynch, 1985; Hacking, 1983, 1988; Heelan, 1983b, 1983c, 1988, 1989a). Where Newtonian positivism holds that metaphysical presuppositions concerning the nature of the thing observed are irrelevant and detrimental to science, scientists and philosophers are coming to terms with the hermeneutic priority of the question, asserting that "the structure of the question is implicit in all experience" (Gadamer, 1989, p. 362) and that "'every sentence I say must be understood not as an affirmation, but as a question'" (Bohr in Holton, 1988, p. 132).

In contrast, the six fundamental philosophical principles Cronbach and Meehl (1955, p. 290-291) offer as support for a "toughminded" set of rules for establishing construct validity begin with the assertion that, "scientifically speaking, to 'make clear what something *is*' means to set forth the laws in which it occurs." These laws may be statistical or deterministic, and are taken to form an interlocking network called the nomological net, pointing toward an identity of what is measured with a theoretical construct. In spite of the apparent definition of construct validity as identity, this identity is never articulated by Cronbach and Meehl "because any suggested identity between constructs and measurements is immediately qualified" (Cherryholmes, 1988, p. 423). The rhetorical qualification of the nomological net's allegedly analytic and logical descriptions and explanations means that the positivist ideal contradicts itself from the start with articulations of the questions, presuppositions, and hypotheses that tacitly structured the observations and made them

meaningful in the first place (Kuhn, 1961, 1970; Latour and Woolgar, 1979; Fahnestock, 1986; Holton, 1988; Polanyi, 1958, 1967)).

The complementary relation between observer and observed that forms the frame of reference in which questions are posed brings a phenomenon into view by delineating the unfolding plot of its story. The signs and symbols expressed in speaking and writing form the basis for an objectivity relevant to both the human and natural sciences. Physicists such as Wheeler (1983), London and Bauer (1983), and Bohr (1983) recognize that science is an activity formed by a community of inquirers, and that this community depends upon its words to hold them together; Bohr emphasized this to the point of stressing that "we are suspended in language such that we cannot say what is up and what is down" (Petersen, 1968, p. 188; Wheeler and Zurek, 1983, p. 5). The process and event of cultural, linguistic, and historical enframing opens up the possibility for creating new meaning, which is why Bohr understood complementarity less as a particular problem associated with quantum mechanics and more as a general epistemological problem of human existence (Holton, 1988, p. 134). In fact, "the most important aspect of the quantum interpretation discussion is the insight it has given into the epistemological role of the conceptual framework" (Petersen, 1968, p. 185). The conceptual framework contextualizes the research question and so determines in advance the direction from which the answer will arrive, doing so, however, without closing off the possibility that the question might be improperly posed, and that indications of an answer might actually be received from another direction (Kuhn, 1961, p. 176; 1970; Gadamer, 1989).

Though Cronbach and Meehl began with largely positivist goals and presuppositions, by the end of the article they explicitly raise issues of interpretation, recommending that the nomological network be altered to cope with anomalous observations (p. 300; Cherryholmes, 1988, p. 424). Where the positivist approach's sense of a completely logical and rational science would deny any role for metaphor (saying one thing but meaning another), for passionate commitment,

or for faith, in establishing the validity of a construct, it is now necessary to admit roles for these factors, so that we may begin the work of focusing our attention on the various objects of the conversation that we are. Thus, "construct validity cannot generally be expressed in the form of a single simple coefficient" (Cronbach and Meehl, 1955, p. 300) because the acceptance of meaning always entails cognitive leaps, or leaps of faith, back and forth between the construct and the measures. These leaps, and the back and forth dialectic between preconceptions of the whole and observation of its parts that they entail, are meaningful insofar as they take up a direction. Gadamer (1989, p. 362) construes the situation to be one in which

> The essence of the question is to have sense. Now sense involves a sense of direction. Hence the sense of the question is the only direction from which the answer can be given if it is to make sense.

To have a direction, questions and answers must converge on some line or train of thought, an "arrow of meaning" (Ricoeur, 1981, p. 193) delineating more and less of the thing in question. Only when they converge, 1) will a clear view of the thing itself emerge; 2) will clarity in thinking be possible; 3) and will the calibration values of items on a test not depend upon the particular respondents; and, conversely, 4) will the measures of the persons not depend upon the particular questions asked in the domain of interest (Rasch, 1960; Wright, 1968, 1977b). It is therefore already evident that sample-free instrument calibration and instrument-free measurement explicitly concern construct validity, which "focuses attention where social theory and research converge and diverge at the juncture of words and things, concepts and objects, theory and practice -- where theoretical constructs and research operations come together and separate" (Cherryholmes, 1988, p. 422). Though data can meet the requirements for measurement modeled by Rasch quite apart from the construct validity of the questions that were asked (Wood, 1978), simply demanding that the questions delineate a direction of more and less leads quite directly to inquiries as to the meaningfulness of that direction, and hence to the investigation of the construct's validity (Wright and Masters, 1982, pp. vi, 12-16, 90-94, 101-106).

One way of getting a handle on the directions in which our various conversations are headed is to look at the kinds of stories that are told by the ways we try to create meaning in life. Cherryholmes (1988, p. 449) uses a passage from Rorty (1985) to contrast traditional and alternative approaches to construct validity. Rorty describes the two principal ways in which people make sense of their lives. In one way, the context in which life is understood is that of historical or fictional heroes and heroines; in the other, life is understood in relation to a non-human, supposedly unchangeable, reality, such as nature. The first way fosters solidarity in community life, the second, objectivity, in the sense of meanings that completely transcend culture and history. Rorty and Cherryholmes stress that the problem with this one-sided sense of objectivity is that it fails to recognize and acknowledge its own cultural and historical embeddedness. I would like to add that the problem with the use of narrative stories in the creation of meaning and validity of constructs is that it fails to recognize and acknowledge its own inherent possibilities for a new, more conversational and playful sort of objectivity.

The Debate. What Jaeger refers to as the Rasch debate is easily construed as a variation on the theme developed by Rorty and Cherryholmes. Jaeger (1987, p. 8) has juxtaposed two quotes that mark the ends of the continuum along which points in the contemporary debate on educational measurement are made:

> There appears to be a fundamental difference in measurement philosophy between those on the two sides of the Rasch debate .... The difference is well characterized in the writings of Benjamin Wright (1968) and E. F. Lindquist (1953). First Wright:
>
> > Science conquers experience by finding the most succinct explanations to which experience can be forced to yield. Progress marches on the invention of simple ways to handle complicated situations. When a person tries to answer a test item the situation is potentially complicated. Many forces influence the outcome - too many to be named in a workable theory of the person's response. To arrive at a workable position, we must invent a simple conception of what we are willing to suppose happens, do our best to write items and test persons *so that their interaction is governed by this conception* and then impose its statistical consequences upon the data to see if the invention can be made useful. (1968, p. 97) [emphasis added; and the quote is actually from Wright, 1977b, p. 97].

In contrast, Lindquist wrote:

> A good educational achievement test must itself define the objective measured. This means that the method of scaling an educational achievement test should not be permitted to determine the content of the test or to alter the definition of objectives implied in the test. From the point of view of the tester, *the definition of the objective is sacrosanct*; he has no business monkeying around with that definition. *The objective is handed down to him* by those agents of society who are responsible for decisions concerning educational objectives, and what the test constructor must do is to attempt to incorporate that definition as clearly and exactly as possible in the examination that he builds. (1953, p. 35) [emphasis added].

Although Jaeger also characterizes the debate as one "between advocates and opponents of the use of IRT [Item Response Theory] in test development and scaling," the debate on the usefulness and meaningfulness of Rasch measurement is conducted as much within what Jaeger would call the IRT community as between it and those outside of it. The debate is therefore taking place on a number of levels, as well as in an international forum. Those advancing various reasons for not using Rasch's approach to educational measurement, or for narrowly restricting its application, include Whitely and Dawis (1974), Whitely (1977), Goldstein (1977, 1979, 1980, 1983), Lord (1980, p. 58; 1983), Divgi (1986, 1989), and Wood (1978). Those rebutting the claims of the critiques include Wright (1968, pp. 99-101; 1977a; 1977b, pp. 102-104; 1984; 1985, pp. 107-109), Andrich (1989), Henning (1989), Gustafsson (1980), Forster (1987), Ingebo (1987), Karr (1987), and Fisher (1991); some Rasch advocates go so far as to suggest that Rasch measurement presents the possibility for a revolution in educational and social measurement (Duncan, 1984; Andrich, 1988; Singleton, 1991; Fisher, 1988, 1989).

Jaeger's quote from Lindquist is a plain and emphatical appeal to a one-sided objectivism in which the discursive activity of construct validation is assumed to take place outside of the context in which the construct is manifest; objections to Rasch measurement almost always hinge on its requirement that objectivity be determined in a less rigid and more flowing manner, from within the interaction of question and answer, as will be shown. Wright, in contrast, is just as

plainly and emphatically struggling with the problem of dealing with the way constructs are simultaneously invented and discovered. Where Lindquist speaks of the sacrosanct, untouchable nature of test items, Wright says that test items amount to nothing more than guesses as to how a construct articulates itself. Wright's suggestion that we observe how well the guesses work to provoke a manifestation of the construct out of the interaction of question and answer, and then see how far the guesses can be made to work in practice, is a fair approximation of what Ricoeur (1981, pp. 212-213; 1976, p. 79) calls the method of converging indices and its probabilistic approach to the validation of guesses. Lindquist wants to disavow the fact that the test items originated in a discursive context, preferring to conceive of them as given in an objective reality. Wright, however, is focusing explicitly on the circular manner in which guesses about reality are entertained, criticized, tested, and applied in an ongoing constructive way.

The extent to which Lindquist is articulating a commonly held position in educational measurement is indicated by the fact that the popularity of multi-parameter models continues despite the fact that they completely defeat any possibility for objective results (Wright, 1984; Andrich, 1988, p. 67), and are difficult and expensive to use (Wright, 1984; Stocking, 1989; Hambleton and Cook, 1977, p. 76; Hambleton and Rogers, 1989, p. 158). The reason for the popularity of two- and three-parameter measurement models in education must be that they will allow the test constructor to accept the validity of test items with no questions asked. On the other hand, the Rasch, or "one-parameter," approach requires the test constructor to pay close attention to the functioning of the items, checking for the extent to which they can be said to hang together along a single continuum of more and less difficulty. The critical evaluation of the performance of the items on the test undercuts the one-sidedness of the test writers' authority by acknowledging the voices of the test takers. Instead of objectifying test takers by subjecting them to an unquestionable authority, the Rasch approach to test construction promotes a conversation in

which questions are tested by the respondents just as much as the respondents are tested by the questions.

And contrary to the theoretical position that, under particular conditions, the results of analyses including two and three parameters are identical to those including only one parameter (Lord, 1980, pp. 189-190), Smith (1990) has quite conclusively shown otherwise. The estimation of the second and third parameters introduces confounding effects even when those parameters take on the values required for fit to an additive conjoint measurement model. Besides, anyone who adheres to rigorous test administration practices will strive to minimize the intrusion of any factors other than the abilities of the persons measured and the difficulties of the problems posed, so it is only reasonable to extend this frame of reference from the data gathering phase of measurement into the data analysis phase (Wright and Stone, 1982, pp. 10-11). This is why Duncan (1984, p. 217), referring to the shared focus of Thurstone and Rasch's approaches to measurement, said that

> 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.

Why allow unexamined presuppositions, prejudices, and preconceptions concerning who the persons measured are and whether the test items actually belong to the same variable to interfere with the measurement process? Should not these be examined, modified, and accounted for, just as much as the students' test behavior and environment is controlled? These questions raise issues best addressed by widening the scope of the debate to include explicit considerations as to what the most important form of test validity is.

Matters of Context. Lindquist is working from within the traditional positivist framework, described by Burtt (1954) as one which defines objectivity as a matter of letting data speak for themselves,

with no recourse to presuppositions or hypotheses allowed. This sense of data arose in historical periods when nature was conceived to be a static constant, with the continents, seas, stars, planets, and biological life precisely the same now as they were on the day God finished the Creation. This sense of data as existing eternally and independent of any human context has fallen under the weight of many different factors, ranging from notions concerning the life cycle of the universe, plate tectonics, and evolution, to the observation that what counts as legitimate data changes from one historical period to another (Kuhn, 1961, 1970; Toulmin, 1953, 1982; Holton, 1988; Hesse, 1970, 1972). However, many of us, like Lindquist, continue to think and act, out of habit, perhaps, as if data are given, not emerging from within a frame of reference.

Messick (1975, p. 959; Cherryholmes, 1988, p. 426) offers a more specific reason for Lindquist's views on educational measurement:

> Construct validity is not usually sought for educational tests, because they are typically already considered to be valid on other grounds, namely, on the grounds of *content* validity. Hambleton and Novick (1973) claim, for example, that 'above all else, a criterion-referenced test must have content validity' (p. 168).

Assuming that tests are valid on grounds of content validity is to be imbued with the overweening confidence that things are as they are because that is the way someone says they are, not because that is the way they actually play themselves out in practice. Empirical examination of the consistency of data may lead to the conclusion that particular test items, and perhaps specific content areas included on a test, may represent constructs different enough in their conceptual structure to invalidate the inferences concerning abilities typically made on the basis of test scores.

The search for construct validity may then contradict the conclusions already drawn concerning the content validity of test items, as Phillips (1986, p. 107) indicates:

> the deletion of misfitting items raises the issue of sacrificing validity for model fit. Typically, achievement test batteries are carefully developed according to detailed content specifications. If items are dropped from a subtest, that subtest no longer matches the test specifications and has lost content validity.

A typical reaction to the suggestion that some items should be deleted from a test is given by Goldstein (1977), who displays the assumption basic to the position that content validity is the only validity relevant to an educational test, namely, that measurement models should be fit to data, in opposition to the Rasch position that data should be fit to a model that clearly specifies criteria for recognizing data good enough to measure with.

> It is by no means clear that the Rasch model does describe real data very well. Willmott & Fowles (1974) admit that when testing the model some items do not fit the model. *These are omitted from the set of items.* As they say, 'The criterion is that items should fit the model, and not that the model should fit the items.' (!) (Goldstein, 1977, p. 310; original emphasis and exclamation; also see Goldstein, 1979, pp. 215-216).

Goldstein (1979, p. 216) is particularly vocal about "moving away from the doctrine of a singly underlying trait, [in order to] allow educational criteria properly to determine test content." But as Gustafsson (1980) points out, items that do not belong to one construct may well belong to another: the problem may be as simple as separately analyzing the groups of items. No one has ever seriously recommended that misfitting items simply be discarded. It is only reasonable to think that items from the same content domain might represent different constructs, and produce data with independent empirical consistencies. The point is to admit that measurement always and everywhere follows from a metaphysics of what counts as an observation (Kuhn, 1961; Heelan, 1972, 1983a, 1985; Heidegger, 1967; Hudson, 1972; Ihde, 1979; Burtt, 1954), and to step into the flow of the hermeneutic circle deliberately and in accord with our intentions.

Focusing on content to the exclusion of the construct re-enacts a fundamental error that has been repeated over and over again in the history of science. The error has makes one of its earliest and most famous appearances in the Pythagorean ontological confusion of representations and images for the things themselves. In the same way that an exclusive focus on content validity precludes attention to constructs,

> Pythagoreans take number and numerical relationships for existence itself and are unable to think of the noetic order of existence by itself, [and so they never] see the real implications of the doctrine of ideas (Gadamer 1980, pp. 35, 32).

The Pythagoreans were caught up in unsolvable problems such as the squaring of the circle and were trying to solve them by means of the physical transcription of the images themselves. Besides forbidding "all recourse and all allusion to manipulations, [and] to physical transformations of figures" Plato redefined the elements of geometry, "denominating such concepts as line, surface, equality, and the similarity of figures" (Ricoeur 1965, p. 202; also see Gadamer 1980, p. 150). Conceiving a point as "'an indivisible line,' and a line as 'length without breadth'" (Cajori 1985, p. 26), Plato construed geometric entities as fictions in order to make the difference between names and concepts as plain as possible; Rasch's (1960, pp. 37-38) comment that "a model is not meant to be true" is intended to have the same effect. The crisis of Pythagorean mathematics was overcome because irrational numbers live out the same conceptual existence in ideality that the rational ones do. The irrationality of the square root of two, for instance, no longer threatened the heart of mathematics because the existence of this number and the line segment it represents no longer depended upon representation as a line segment of precisely drawable length or as a number that could be exactly specified. The crisis of educational measurement provoking the Rasch debate hinges on the same problem, namely that the rationality of testing procedures depends on whether what is measured is denoted by content (name) or construct (concept).

The point in using figures of any kind, whether they are metaphorical, numerical or geometrical, is to facilitate clarity in thinking through clear representation of the thing itself. Clear views on entities are brought about when it is possible to look through the content of the particular figure drawn and see the thing that remains constant free of influence from the particular representation. Plato's restrictions on the use of instruments in geometry to the compass and straightedge was aimed at the possibility of allowing things to communicate themselves, not by confusing the conceptual ideality of things with their names, as Pythagoreans and positivists do, but by using the instruments as channels or media for the expression of the things themselves. Plato placed philosophy in close association with mathematics because

> even he who has not yet seen all the metaphysical implications of the concept of pure thinking but only grasps something of mathematics . . . knows that in a manner of speaking one looks right through the drawn circle and keeps the pure thought of the circle in mind (Gadamer 1980, p. 101).

Gadamer could easily be paraphrasing Plato In Book VI of the Republic (510d) Plato writes that mathematicians

> make use of the visible forms and talk about them, though they are not thinking of them, but of those things of which they are a likeness, pursuing their inquiry for the sake of the square as such and the diagonal as such, and not for the sake of the image of it which they draw.

Such is not the case, though, when mechanical devices for copying and reproducing angles and curves are introduced with the aim of solving problems such as the squaring of the circle, as was often done in ancient Greece (Cajori 1985, p. 27; Bunt, Jones and Bedient 1976, p. 126; Ball 1919, pp. 28, 35, 43). In these instances, things do not communicate themselves, and thinking is not an element of being, because concept is confused with name and thinking presumes to precede and communicate being.

But geometrical analyses are not valid just because they are performed on geometrical figures such as circles and triangles; beyond the validity of the content, it is essential to establish the validity of the construct, to distinguish between the content of the items and the validity of taking them as representative of a conceptual dimension. "Since predictive, concurrent, and content validities are all essentially *ad hoc*, construct validity is the whole of validity from a scientific point of view" (Loevinger, 1957, p. 636 in Messick, 1975, p. 956), and so, "*all measurement should be construct-referenced*" (Messick, 1975, p. 957). Loevinger's (1965, p. 151) appreciation for Rasch measurement cannot be separated from her position on construct validity, since "any concept of validity of measurement must include reference to empirical consistency" (Messick, 1975, p. 960). Whitely, on the other hand, holds to the explicitly positivist end of Cronbach and Meehl's (1955) sense of construct validity as "appealing to criteria outside of the measuring

process ... in accordance with a nomothetic network" (Whitely, 1977, p. 232), which is exactly the way Lindquist and Goldstein see the matter.

Wright's (Wright and Masters, 1982, p. 91) concept of construct validity is much closer to Cherryholmes, Loevinger, and Messick's discursive formulation than it is to Whitely's positivist construal:

> The responses of each person can be examined for their consistency with the idea of a single dimension along which items have a unique order. Unless the responses of a person are in general agreement with the ordering of items implied by the majority of persons, the validity of the person's measure is suspect.

The same dialectical relation between whole and part holds for items:

> Responses to each item must be examined for their consistency with the idea of a single dimension along which persons have a unique order. Unless the responses to an item are in general agreement with the ordering of persons implied by the majority of items, the validity of the item is suspect.

The history of science concurs with the discursive formulation of construct validity and disputes the positivist because of the crucial importance of the ontological difference between mathematical and perceptible being. This difference is what "Eudemos singles out [as] Plato's contribution in his history of mathematics, namely, to have distinguished between name and concept (Simplicius Physics 98)" (Gadamer 1980, p. 100). In the same way that Plato resolved the Pythagorean over-complications with mathematical clarity and simplicity, Copernicus, Kepler, and Galileo founded modern science when they resolved the Aristotelian astronomical complications by basing their studies on mathematical idealizations and observations. Cronbach and Meehl (1955) focused attention on the difference between content and construct, and brought social measurement a step nearer to recovering the meaning of mathematical clarity, but Rasch's restrictions on measuring instruments have the potential of re-creating in contemporary social science what Plato's and Galileo's restrictions on and uses of measuring instruments did for geometry and natural science. If this potential is realized, claims concerning the revolutionary status of the Rasch models (Andrich, 1988; Loevinger, 1965; Singleton, 1991) will be justified.

Jaeger's Revolution Revisited. Jaeger (1987) juxtaposes the quotes from Lindquist and Wright in the context of alternately proclaiming and questioning the revolutionary status of developments in educational measurement over the last twenty years. Just as Wright (1984, for example) often does, Jaeger (1987, pp. 9-12) uses quotes from Thorndike and Thurstone as evidence of the age and importance of some of the most fundamental ideas in educational measurement. But Jaeger does not seem to realize that the revolution in educational measurement begun by Thorndike, Thurstone, and others is still happening; he certainly gives no hint as to what the point of the revolution might be. The contextual matters crucial to understanding the Rasch debate have provided some clues as to what that point might be. Kuhn (1970) suggests more to look for when he indicates that observational anomalies, methodological problems in accounting for them, and resulting degrees of extreme complication prepare the ground for scientific revolutions. Thus, the Pythagorean and Aristotelian over-complications and rationalizations that Plato and Galileo cut through with their insistence on rigorous observation and mathematical idealization in the use of the compass, straightedge, and telescope may have their parallels in the two- and three-parameter IRT models and the fixation on content validity plaguing educational measurement.

Just how complicated has the focus on content validity made measurement for IRT? That IRT's stress on the two- and three-parameter models follows from Lindquist's demand that analysts not "monkey around" with item content is evident in Hambleton and Novick's (1973; quoted by Messick (1975, p. 959)) opinion that content validity is important "above all else" in the construction of tests. Is it any wonder, then, that Hambleton (Hambleton, 1983; Hambleton and Cook, 1977; Hambleton and Rogers, 1989) and Novick (Lord and Novick, 1968) are two of the names most immediately connected with IRT's stress on the two- and three-parameter models? These models make no recommendations as to potential threats to construct validity because content validity is the matter of interest. The second and third parameters do little more than act to allow the inclusion of just about any data in an analysis, as is implied by Whitely's (1977, p. 229) comment

that "the several studies which apply a reasonably stringent test of fit are notable for the frequency with which the [Rasch] model is found to be inappropriate". Goldman and Raju (1986, p. 19) concur, saying that since the findings of their "study suggest that the two-parameter model fits the attitude survey [of interest] better than the Rasch model, future applications might emphasize the two-parameter model." Hambleton and Rogers (1989, p. 148) are direct, saying that "the one-parameter model has rarely provided a satisfactory fit to the test data; the three-parameter nearly always has." Given the successful fit of test data to the Rasch models that is commonly achieved in the calibration of the widely-used Rasch-based item banks (Wright and Bell, 1984), one can only wonder just how poor the data fitting the three-parameter model are.

What can the significance of the observations made by Whitely, Goldman and Raju, and Hambleton and Rogers be, except that retaining items on a test is more important than understanding the variable, and that content is more important than assuring ourselves that comparisons of more and less are comparisons of more and less of the same thing? Divgi (1986, p. 283) says as much: "Issues like 'objectivity' and consistent estimation are shown to be unimportant in selection of a latent trait model." The stress on content validity has led to a situation in which tests teach us nothing more than what happened when a particular group of students responded to a particular group of items. Would not it be far more productive to investigate the structure of constructs, assign measures, and calibrate instruments free of concern (within the specified frame of reference) for the particular persons and items involved? Pretensions as to the sample-free measurement properties of IRT and its "strong" assumptions concerning unidimensionality are belied by simple mathematics (Wright, 1985, pp. 107-109; Fisher, 1991) and by a desire to retain test items that can be satisfied only by models designed to fit almost any data.

More echoes of Lindquist's appeal to the authorities on high, the sacrosanct nature of test items, and the prohibition against monkeying around with item content resound when Messick (1975, p. 959) quotes Osburn (1968), who says that

> what the test is measuring is operationally defined by the universe of content as embodied in the item generating rules. No recourse to response-inferred concepts such as construct validity, predictive validity, underlying factor structure or latent variables is necessary to answer this vital question (p. 101).

Cherryholmes (1988, p. 453) observes that this sort of ultra-operationalism had been rejected even by the logical positivists more than 30 years before Osburn wrote, because they saw that the cognitive significance of concepts is never generated in a strictly rule-following way. Cronbach and Meehl accordingly rejected operationalist definitions of constructs from the beginning of their study of construct validity.

The importance placed on test content in education received subtle and implicit support from those who objected to Stevens' (1946) proscriptions on the nature of scales and the statistical operations appropriate to each. Among those objecting was Lord, who said that "the numbers don't remember where they came from," meaning that no unscientific consequences arise from performing arithmetic operations on any kind of numbers, no matter if they are nominal, ordinal, interval or ratio (Lord, 1953, p. 751). By 1980, however, Lord had apparently forgotten that the means and standard deviations of nominal "football numbers" "will obey the usual mathematical relations that have been proven to be applicable to random samples from any normal population" (Lord, 1953, p. 751). In his effort to justify the discrimination and guessing parameters in the IRT models, Lord (1980, p. 58) asserted that "there is no sufficient statistic when the item response function is a normal ogive, even though the normal ogive and logistic functions are empirically almost indistinguishable," the point being that if there is no sufficient statistic in the first place, then it is no use trying to set and meet standards of additive conjoint measurement, such as those offered by Rasch (Perline, Wright, and Wainer, 1979; Brogden, 1977). However, no practical difference results from the use of the normal ogive (Cohen, 1979; Wright and Stone, 1979, pp. 60-61; Wright and Masters, 1982, pp. 77-80). Even though his reasons for doing so are exactly opposite Lord's, Wright (Wright and Stone, 1979, p. 130) uses Tchebycheff's inequality to justify

the use of the normal ogive, just as Lord (1953, p. 751) does in his argument concerning arithmetic manipulations of nominal "football numbers."

Lord (1980, p. 5P) takes his argument against the usefulness and meaningfulness of setting high standards for data quality a step further, saying that the presence of guessing also defeats the possibility for sufficient statistics. Because guessing presumably is an ineradicable possibility in educational measurement, the focus on content validity is retained and the implications for construct validity raised by taking sufficient statistics as a heuristic goal are ignored. Within the framework of this rationalization, Lord is able to overlook basic points in the mathematics needed for establishing unidimensional constructs, presumably because "the numbers don't remember where they came from." The Rasch model is thereby construed to be a special case of the two- and three-parameter models, and is effectively "included in the three-parameter model" (Lord, 1980, p. 190) because the discrimination and guessing parameters default to the values "assumed" by the Rasch models when data in fact live up to the requirements of measurement.

Lord is mistaken about this and not only because Smith (1990) has shown large and important differences between three-parameter and Rasch analyses. More important are the mathematical connections that have been established showing the necessity and sufficiency of Rasch's models for objectivity in measurement (Rasch, 1960, 1961; Douglas and Wright, 1986; Wright and Douglas, 1986; Andersen, 1973), and their embodiment of the principles of measurement delineated by Campbell (1920), Thurstone (1928), Guttman (1950), and Luce and Tukey (1964), as has been shown by Perline, Wright, and Wainer (1979), Brogden (1977), Brink (1972), Andrich (1978, 1985), Englehard (1984, 1989), Jansen, (1984), Jaeger (1987, pp. 9-11), and Wright (1985, 1988, 1989; Wright and Masters, 1982, pp. 1-10; Wright and Stone, 1979, pp. vii-x), none of which is possible for the two- and three-parameter models. This would not seem to be too important to IRT advocates, however, since "it will be a sad day indeed when our conception

of measurable educational achievement narrows to the point where it coincides with the criterion of fit to a unidimensional item response model" (Traub, in Hambleton, 1983, p. 64).

Given this state of affairs it is ironic that "the most important IRT assumption that the commonly used models must satisfy is unidimensionality, an assumption requiring that the test items measure just one underlying ability dimension (Lord, 1980)" (Phillips and Mehrens, 1987, p. 1). The situation is more than just ironic, though, when one reads about "one-dimensional IRT models (with 2 or 3 parameters)" (Hambleton and Rogers, 1989, p. 158). The potential for unidimensionality in IRT models is destroyed by the inclusion of the discrimination and guessing parameters. The comparison of abilities is scientific only when the same magnitude of difference results no matter which particular item or instrument of the kind in question is used. Allowing extra parameters to vary across the items means that the difference between two abilities will inherently depend upon which item is used to make the comparison. This fact is a simple mathematical relation, but does not prevent some IRT advocates from nonetheless making misguided claims to sample-free measurement, probably because of the persistence of the illusion that Rasch measurement is a special case of the three-parameter model.

As might be expected from item response models whose estimation algorithms contradict their own "assumptions" of unidimensionality, the most commonly used computer program for implementing the two- and three-parameter IRT models, LOGIST (Wingersky, Barton, and Lord, 1982), has been shown by Stocking (1989, p. 42) to be rife with "large (and sometimes unacceptable) biases" in the estimation of the parameters. Stocking took up the study of LOGIST-based applications of IRT in order "to explore and understand some apparently anomalous results ... that have been obtained from time to time over the past several years" not only in real data, but also in data simulated to fit the three-parameter model. After remarking, in a manner reminiscent of many of her colleagues (documented in Wright, 1984), on the expense and difficulty of using LOGIST, Stocking (1989, pp. 44-45) concludes that

> LOGIST ... needs improvement. Most applications cannot afford to run the program to complete convergence. It may be possible to improve results of the four-step structure by obtaining better starting values for the parameters. Alternatively, controlling the behavior of estimates of discrimination and guessing parameters through the imposition of prior distributions on them may be cost effective and provide reasonable results.

The four-step procedure (Stocking, 1989, p. 21) referred to is one in which abilities and difficulties are estimated first, holding the discrimination and guessing parameters constant; then, the abilities are fixed and the three item parameters are estimated. Steps three and four repeat the first two steps. This structure was imposed on the estimation procedure in an effort aimed at overcoming the tendency of parameter estimates to diverge without limit (Stocking, 1989, pp. 25-26). Lord noted quite some time ago that "the method usually does not converge properly" (Lord, 1968, p. 1015) and that "experience has shown that if ... restraints are not imposed, the estimated value of [discrimination] is likely to increase without limit" (Lord, 1975, p. 14). These problems are precisely what caused Wright to reject the multi-parameter approaches in the mid-1960s, when he and Bruce Choppin wrote such programs against Rasch's advice (Wright, 1988, p. 3). LOGIST's four-step procedure uses the Rasch model, in effect, every other iteration through the data (on the first and third steps of the four-step procedure) in order to provide "reasonable estimates for item parameters and abilities in a feasible amount of time" (Stocking, 1989, p. 21).

Stocking (1989, p. 45) indicates that these complications are not simply a matter particular to LOGIST when she makes the same recommendations concerning another program, BILOG (Mislevy and Bock, 1983):

> BILOG, being a more recent computer program available for general use, has not been subjected to the same wide variety of applications as LOGIST. As such, it does not contain the necessary restrictions to prevent the numerical procedures from diverging from reasonable, although perhaps less than optimal starting values. It seems clear that such additional restrictions are necessary.

"Better starting values for the parameters," and "imposing prior distributions on them," are "necessary restrictions" that the two most widely used IRT computer programs must incorporate just to provide "reasonable estimates ... in a feasible amount of time." Wright (1988, p. 3) realized

the same thing about his own two-parameter program in 1964, saying that it would not "converge unless I introduced some inevitably arbitrary constraint. The choice of the constraint would always alter the results .... Since I couldn't make the two-parameter program work, I discarded it." Hambleton and Rogers (1989, p. 158) also comment on the unavailability, unfriendliness, cryptic and unwanted output, and bugs of IRT computer programs, in addition to the excessive time and prohibitive sample sizes required for their application. In contrast, Hambleton and Cook (1977, p. 88) write that "the problem of ability and item parameter estimation with the Rasch model is quite different. In fact, the estimation problem is essentially resolved." It is also interesting to note the continued relevance of Hambleton and Cook's (1977, p. 76) comment that the only "fast and convenient-to-use computer programs for estimating the parameters [are those available] for the Rasch model." Wright (1984) documents more words of praise, from those who primarily advocate two- and three-parameter models, for the efficiency and effectiveness of Rasch's approach to measurement. Because the two- and three-parameter models often will not work at all with small sample sizes, Lord (1983) has said that small sample sizes justify the use of the Rasch model. The principles of additive conjoint measurement would then appear to be the best route to take for the great majority of tests, since most of these are administered in classrooms with less than fifty students.

The willingness to go on struggling against the excessive expense and complication of IRT apparently has deep roots in the traditional educational emphasis on content validity. But the focus on content turns educational tests into devices akin to the mechanical devices that Plato excluded from geometry: devices that confuse perceptible being for the thing itself, and introduce over-complications which cut off sight of the pure being of geometrical figures by purporting to communicate a thing better than it could communicate itself. The impossibility of squaring a circle or of duplicating a cube evidently gave rise to many attempts at mechanical solutions. Trying to create fair and scientifically valid tests by focusing on the content of the items amounts to exactly

the same sort of impossible problem solvable only by means of sophistic manipulations. But in these attempts, Plato claimed, "the good of geometry is set aside and destroyed, for we again reduce it to the world of sense, instead of elevating and imbuing it with the eternal and incorporeal images of thought" (as quoted by Cajori 1985, p. 27). Though the eternity of these images is doubtful, their incorporeality constitutes the spirituality and ideality of language, and this is what is stifled and ignored by educational testing's preoccupation with content validity.

It does appear that the most important aspect of validity in American educational measurement is the capacity to tell what Rorty (1985) calls stories of objectivity, in the sense that objectivity is the one-sided imposition of authority. Most educational measurement experts are willing to allow issues of construct validity to be decided by default, and "if researcher-theorists default on construct validity, then they consciously or unconsciously adopt inherited discourses and meanings previously assigned to constructs and measurements" (Cherryholmes, 1988, p. 428; also see Gould, 1981). As Burtt (1954, p. 225) phrased it,

> What kind of metaphysics are you likely to cherish when you sturdily suppose yourself to be free of the abomination? Of course . . . in this case your metaphysics will be held uncritically because it is unconscious; moreover, it will be passed on to others far more readily than your other notions inasmuch as it will be propagated by insinuation rather than by direct argument.

The positivist denial of metaphysics is also assumed anytime someone purports to be able to count on test items to provide valid and reliable measures when no value is placed on checking whether it is reasonable to add up counts of right answers and assign scores. However, just because experts have decided that items on a test all belong to the same content domain does not mean that they belong to the same construct.

Viewed in this larger context, what Jaeger (1987) called "the Rasch debate" begins to look more like "the validity debate." Content validity is far more of an ideological, bureaucratic, and administrative need than a scientific or a human one. Many observers have suggested that educational measurement addresses the social, economic and political agenda of elite decision-

makers more than it does the interests of equal opportunity and justice (Crouse and Trusheim, 1988: Owen, 1985; Sutherland, 1984; Strenio, 1981), and it will continue to do so until more attention is paid to the discourse processes and metaphysics of testing. Cherryholmes (1988, p. 421) suggests that some attention to these issues began, and "social research methodology entered adolescence, if not maturity, in July 1955 ... with the publication of Cronbach and Meehl's 'Construct Validity in Psychological Tests." The problem is that "the adolescence has been arrested" (Cherryholmes, 1988, p. 450). If so, the potential for its further development grew with the publication of Rasch's research on measurement (1960), as has been noted by Duncan (1984). That potential will hardly begin to be realized until educators overcome their fixation on content validity, however.

Implications for Practice. Sensitivity to the role of culture in the work invested in the framing of questions has led to a new emphasis on qualitative, ethnographic style research in education. Though this development has been productive in promoting a more dialectical critique of the question and answer process, little or nothing in the way of suggestions for improvements in quantitative thinking have been forthcoming; quar. 'ative methods have been either relegated to the positivist trash heap of history by qualitative purists, or accepted as unavoidably positivist, at least in part, by most of those who still continue to use and think about them. Even those who recognize the philosophical problems attending quantitative methods and incorporate a critical dialectic into their application, such as Cook and Campbell (1979, pp. 91-94), still take only roundabout routes at best to show that their data are focused on a common question, point in the same direction, and that the responses received have arrived from that direction.

A more direct approach is to specify what will count as an observation in advance, on the basis of informal observations, hunches, or previous research, focus questions on the continuum along which the variable will likely be manifest, and examine the questions for conformity to measurement principles after they have been exposed to a relevant group of persons (Rasch,

1960; Wright, 1968, 1977b). Where education's traditional concern with content validity moves straight from the theoretical construct to observation to assertions concerning what is observed (Cherryholmes, 1988, p. 448) in a monological and one-sided fashion, Rasch and Wright insist on the importance of completing several spirals through the hermeneutic circle, returning to check and possibly alter observations and theoretical constructs before making hard and fast assertions about what has been observed or what can be expected in the way of future observations. Cherryholmes (1988, p. 448; also see Fisher, 1988b, 1989, 1990) says that "quantitative and qualitative approaches are combined when the meaning of these bidirectional arrows [moving from construct to observation to phenomenon and back again] is clarified and negotiated." What Cherryholmes (1988, p. 448) refers to as the "'covariation' or shared meaning but not identity" connoted by these arrows has been called a "mutually critical correlation" in the methods of theology (Tracy, 1975), and traces a dialectical spiral that delineates the "arrow of meaning" followed in conversation or the reading of a text (Ricoeur, 1981, p. 193). The same mutual relation of construct to phenomenon, mediated by the structure of a text, occurs when data meet the requirements of additive conjoint measurement. Focusing the research question by attending to the ways in which it is posed by the test or survey questions extends and refines the question and answer process by which meaning is created in conversation, or by which meaning emerges from the reading of a text.[1]

The criticism of quantitative methods must be complemented by criticism of qualitative and ethnographic approaches that emphasize only the return arc of the dialectic, which makes them

---

[1] Unfortunately, after Cronbach (1982, p. 70) came to appreciate the rhetorical and qualitative implications of belief in the construal of reality, he considered Rasch (1961) to hold that "one-parameter scaling can discover coherent variables independent of culture and population." On the contrary, Wright himself could have written what Cronbach says on the next page, that "the sooner all social scientists are aware that data never speak for themselves, that without a carefully framed statement of boundary conditions generalizations are misleading or trivially vague, and that forecasts depend on substantive conjectures, the sooner will social science be consistently a source of enlightenment."

just as incomplete as the quantitative approaches that follow only its leading arc. Neither approach alone successfully addresses the problem of method in social research, and their simple side-by-side juxtaposition does not accomplish anything of substance, either. What is required is a more fully complementary relation between the two, wherein each subtly incorporates what is most important about the other into its own movement, acknowledging in practice that "the social roots of social measurement are in the social process itself" and that "quantification is implicit ... in the social process itself before any social scientist intrudes" (Duncan, 1984, pp. 221, 36). Only then does the phenomenologically rich sense of method as the activity of the thing itself come into play (Gadamer, 1989), and this is precisely what Rasch offers. The activity of the phenomenon measured moves first in the direction shared by the questions on a test toward the responses they provoke; the responses in turn raise new questions which either extend or otherwise alter the direction initially followed. The back and forth motion continues in a manner that connects with what is most fundamental to method, the way in which clear thinking follows after and is carried along by (*meta-hodos*) the thing itself along the path of meaning it cuts within a particular cultural and historical frame of reference. This is not to say, however, that additive conjoint measurement models embody the essence of method, or that they even are methods, because they are not. The methods by which meaning is created vary substantially both among and within areas of interest. The point is only that education's obsession with content validity tries to cut off the flow of method prematurely, meaning that a shift in focus toward construct validity would contribute to the phenomenological and methodological soundness of educational research.

The recent surge of interest in fit analysis, differential item functioning, and the Mantel-Haenszel procedure is a move in the direction of a stronger emphasis on construct validity, but still presumes an approach to measurement often lacking in the methods creating the data to which it is applied. For instance, in the application of the Mantel-Haenszel procedure,

> If one is not prepared to accept the validity of the Rasch model for the item under examination, the implicit assumptions of the MH procedure will not be satisfied either. If one is prepared to accept the Rasch assumptions, however, the Rasch model yields simpler and better statistics (Linacre and Wright, 1987, p. 16; 1989, p. 3).

Thus, the application of the MH procedure to data that fit the three-parameter IRT model but not the Rasch model adds yet another level of self-contradiction and complication to educational measurement. The residual differences between modeled and observed responses calculated by both the Rasch and the MH procedures implement the rigorous sense of unidimensionality contradicted by the two- and three-parameter estimation algorithms. What is the point of obtaining more complex and hard to interpret statistics from the MH procedure when a model that almost always fits data is being used to provide ability and difficulty estimates? Why not use the same requirements used to calculate fit to estimate scale positions, and arrive at simpler statistics in less time and with less trouble? Could it be that preconceptions concerning the items' inherent content validity make it easy for educational measurement specialists to continually find new complications that prevent or dislocate a focus on construct validity?

The sort of structure required of data for fit to an additive conjoint measurement model, and presumed in the application of the MH procedure, is displayed in Table 1. In fact, it is only reasonable to count up marks of correct and incorrect (or marks of correct, partly correct, and incorrect (see Masters, 1982, for more on partial credit scoring)), and use the counts as a basis for making inferences about person ability or item difficulty, when data can be organized into a pattern roughly similar to the one shown in Table 1. The items are ordered from more to less difficulty according to the number of persons responding correctly to each; the persons are ordered from more to less ability according to the number of items to which each has correctly responded. The resulting pattern required for measurement is one in which a person may occasionally score a correct answer after missing an item or two, but there is a general harmony to the continuum of more and less shared by the persons and items.

In contrast, Table 2 displays data that contradict the basic requirement of unidimensionality, and so threaten the construct validity of the calibrations and measures. Imagine that the data in Table 2 are embedded in a large matrix of data organized like that shown in Table 1, in which a general order of more and less of something remains relatively and probabilistically constant across items and persons. Every person in Table 2 has the same count of correct answers, but is it possible to assume that the counts mean the same thing? Is not that assumption made, however, every time a teacher or a tester computes the percentage of the total number of items to which a student responded correctly? In contrast to Divgi (1986, p. 283), Messick's (1975, p. 960) answer to this question is an unequivocal yes:

> Inferences in educational and psychological measurement are made from scores, and scores are a function of subject responses. Any concept of validity of measurement must include reference to empirical consistency. Content coverage is an important consideration in test construction and interpretation, to be sure, but in itself does not provide validity.

After all, is not it possible that some students will respond to ostensibly easy questions incorrectly, and ostensibly hard ones correctly, independent of the fact that all of the items belong to the same content domain? Is not it important to detect when this sort of thing happens on a large scale, as has been the case with Anne, Igor, Larry, and especially Joe, in Table 2? And what about Bob, who was correct on every other item when they are ordered by difficulty? Is he making some kind of joke? The probability of Igor missing the easiest item must be very small, so was this the result of simple carelessness or is something more important going on? Anne and Larry both got the very hardest item correct after missing five in a row. Is this simply a sign of some special knowledge they each have, or did they collaborate on the answer? Answers to these questions can be gained by asking the students new questions similar in difficulty to those on which they have provided surprising responses. If the items in Table 2 are in entry, as well as measure, order, perhaps it would be beneficial to ask if Mary ran out of time as she labored with each question before she moved on to the next. Did Joe perhaps skip all of the easy questions out of boredom?

Did Bob make random marks on the answer sheet? If so, why? Will Larry and Anne answer another item of question 10's difficulty correctly, or were their responses produced by cheating, guessing, or special knowledge? Would Igor have missed the first question if he had not been in such a hurry to get started, or if he had not had difficulty figuring out what the test was about?

These examples are intended to show that there are many kinds of disturbance that interfere with the effort to measure, and each is as likely to occur as guessing is, and will present just as much potential for disruption. Are we to then incorporate additional parameters for plodding, sleeping, and fumbling, as they are called by Wright and Stone (1979, pp. 170-190)? Hardly; two basic reasons for the movement toward qualitative methods in educational research is that usual applications of quantitative method traditionally strive to anticipate, close off, trap, or nail down anomalies, and to focus on operations and content instead of meaning and constructs. It is more sensible, though, to go with the flow of the multi-faceted, conversational and metaphorical logic by which things actually play themselves out, than it is to force a one-sided logic and rationality on what people do. Well put questions inevitably open up more questions than they answer, and to cut off questioning is to kill the potential for learning. Disruptions in the measurement process are inevitable but it is far more productive to locate and interpret them after they occur than to try to include them as elements in a model of an already very complicated situation.

Patterns of anomalous response commonly found in educational test data are discussed in Wright and Stone (1979, pp. 170-190). Quantitative methods for flagging unexpected patterns of response associated with persons and items are standard equipment in programmatic applications of the Rasch models, such as BIGSCALE (Wright, Linacre, and Schultz, 1990). More complex multiple regression procedures using the conceptual structure of item characteristics to predict Rasch item difficulties have been presented by Stenner and Smith (1982) and Stenner, Smith, and Burdick (1983) in the context of exploring construct validity.

The interpretive study of an ordered data matrix shows that scores are meaningful only within the context of a frame of reference, and that the Rasch model's requirement of shared order across persons and items is in fact assumed whenever raw scores are used as a basis for comparison, Goldstein's (1979, p. 219) claims to the contrary notwithstanding. In Wright's (1977b, p. 114; also see 1985, pp. 106-107) terms,

> Unweighted scores are appropriate for person measurement if and only if what happens when a person responds to an item can be usefully approximated by the Rasch model.... Ironically, for anyone who claims skepticism 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.

Empirical studies completed on many different kinds of test, survey, and rating scale data have answered the question concerning the Rasch model's practical usefulness in the affirmative many times over, as is evidenced by just a cursory examination of the papers presented to the Midwest Objective Measurement Seminars, the International Objective Measurement Workshops, and the Rasch Measurement SIG sessions of the AERA, besides the publications appearing in journals as diverse as the *Archives of Physical Medicine and Rehabilitation* and the *Journal of Coatings Technology*. The medical fields have found Rasch's approach to measurement especially useful, with a great deal of Rasch applications being found in accreditation and certification, besides in psychiatry, nursing, and blind and physical rehabilitation.

One of the most important ways Rasch measurement will influence the future of education is in the convergence of Vygotsky's (1978) notion of the zone of proximal development with the concept of measure-driven instruction, or curriculum-referenced measurement (Rentz and Bashaw, 1977, Wright and Bell, 1984). Some presentations of measure-driven instruction focus on "high-stakes" tests, the ones that must be passed for graduation or certification (Airasian, 1988), but the concept is more appropriately placed in the context of day-to-day instruction, where the anxiety associated with tests will be reduced and defused instead of blown even more out of proportion.

Of course, measure-driven instruction of any kind is doomed to fail if the tests are validated by content and not by construct because teaching to the test becomes a problem only when specific content areas not conceptually related to the structure of the test are known to be included. On tests that adhere strictly to the principles of measurement, special content areas have no place; experience has shown that content-related issues introduce negligible or easily managed amounts of disturbance on tests constructed from a pool of items calibrated and validated for representation of the construct.

The connection with the zone of proximal development comes with the realization that well-constructed tests are targeted at the ability of the person measured. For instance, Table 1 could be used to inform learning objectives for the students' measured. The lessons following the test could be individually tailored for each student to start with a review of the topics represented by the hardest items answered correctly, and moving from there into new territory with the topics represented by the easiest of the items answered incorrectly. This strategy would probably be best administered in conjunction with computerized adaptive tests capable of pinpointing abilities with few numbers of items. The frequent taking of tests and the practical application of their results would go a long way toward making testing more a matter of measurement than one of anxiety for the test taker.

Conclusions. In contrast to the way Rorty and Cherryholmes put it, I would like to stress the fact that stories of solidarity and objectivity are not mutually exclusive. Cherryholmes (1988, p. 450) goes only halfway towards making this point:

> If Rorty is correct that reflective human beings make sense of their lives by telling stories about either solidarity or objectivity and our stories about objectivity are flawed, they nevertheless describe a community. The community is elitist, control centralized; criticism is limited to experts; the social context and historical setting of the community is not discussed; constructs (the way the community is conceptually organized) are not chosen on ethico-political or aesthetic grounds but in terms of 'scientific' criteria; and the discourse is thought of as nonmaterial and descriptive-explanatory.

To this it must be added that, if the solidarity of societies emphasizing objectivity is likely to take a one-sided, dictatorial, and authoritarian form, then the objectivity of societies that emphasize solidarity is likely to be two-sided, conversational, and playful (Heelan, 1983a, 1985, 1989b; Ihde, 1979; Ackermann, 1985). There is a large literature describing scientific processes and results in the language of community life (Ormiston & Sassower, 1989; Fleck, 1979; Hesse, 1970, 1972; Toulmin, 1953, 1982; Latour & Woolgar, 1979; Holton, 1988; Kuhn, 1961, 1970); the problem these works address is how to find and nurture whatever resources for solidarity we may have remaining to us in our scientific society. In no way does this require us to abandon objectivity or severely delimit its sphere; on the contrary, we aim to avoid yet another simplistic reduction of life's richness to yet another mere dichotomy.

In opposition to Lindquist's approach to measurement, Wright specifically addresses ethical, political and aesthetic criteria by which to judge and choose constructs. Because we intend to use our measures to inform decisions that affect people's lives, we are ethically bound to be sure that the numbers actually represent more and less of the construct in question. The only ethics addressed by Lindquist concern a blind devotion to following orders. Because we are legally and morally bound not to discriminate among persons by religion, sex, race, sexual orientation, or age, we require that our measures not vary across these groups in an inordinate fashion. Lindquist's definition of the construct as sacrosanct prevents attention from being focused on these issues in an effective way. Rasch's measurement models offer an aesthetically pleasing symmetry of question and answer in which each plays itself out in terms of the other, effectively extending and furthering the process by which meaning is reproduced in social life, conversationally. Lindquist, on the other hand, would have us only accept that which is handed down without question because we have no business monkeying around with sacrosanct definitions.

Educational Import. The desire to understand human experience by means of stories taken from a non-human, ahistorical reality still predominates in much of social science. In education this

desire is evident in the popularity of measurement models that do not recognize or accept the fact of their own imposition of political, moral and aesthetic criteria upon test items and data. By recognizing the inevitability of such impositions and formulating models of how these criteria can be simply, easily and practically implemented, explicated and criticized, Rasch and Wright have made an enormous contribution to the project of a scientific humanity.

TABLE 1

SAMPLE DATA THAT DISPLAY THE RECIPROCAL ORDER NEEDED FOR CONVERGENCE AND FIT TO AN ADDITIVE CONJOINT MEASUREMENT MODEL

	Items Easy or agreeable to hard or disagreeable										
Persons	1	2	3	4	5	6	7	8	9	10	Person Scores
Luc . . .	0	1	0	0	0	0	0	0	0	0	1
John . .	1	0	1	0	0	0	0	0	0	0	2
Louise . .	1	1	0	1	0	0	0	0	0	0	3
Martha . .	1	1	1	0	1	0	0	0	0	0	4
Jimi . .	1	1	1	1	0	0	1	0	0	0	5
Diane . .	1	1	1	1	1	1	0	0	0	0	6
Nathan . .	1	1	1	1	1	1	0	1	0	0	7
Jon . . .	1	1	1	1	1	1	1	1	0	0	8
Laura . .	1	1	1	1	1	1	1	0	1	1	9
Alissa . .	1	1	1	1	1	1	1	1	1	0	9
Item Score	9	9	8	7	6	5	4	3	2	1	

TABLE 2

SAMPLE DATA ON THE VARIATION OF MEANING IN A SCORE

	Items Easy or agreeable to hard or disagreeable										
Persons	1	2	3	4	5	6	7	8	9	10	Person Scores
Joe	0	0	0	0	0	1	1	1	1	1	5
Mary	1	1	1	1	1	0	0	0	0	0	5
Lucy	1	1	1	1	0	1	0	0	0	0	5
Bob	1	0	1	0	1	0	1	0	1	0	5
Anne	1	1	1	1	0	0	0	0	0	1	5
Larry	1	1	1	1	0	0	0	0	0	1	5
Igor	0	1	1	1	1	0	1	0	0	0	5

References.

Ackermann, John Robert. (1985), Data, Instruments, and Theory: A Dialectical Approach to Understanding Science. Princeton: Princeton University Press.

Airasian, Peter W. (1988), "Measurement driven instruction: A closer look", Educational Measurement: Issues and Practice, 7(4), 6-11.

Andersen, E. B. (1973), "A goodness of fit test for the Rasch model", Psychometrika, 38(1), 123-140.

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

_____. (1985), "An elaboration of Guttman scaling with Rasch models for measurement. In Sociological Methodology 1985. Ed. N. B. Tuma. San Francisco: Jossey-Bass.

_____. (1988), "Educational and other social science measurement: A Kuhnian revolution in progress", Unpublished ms.

_____. (1989), "Statistical reasoning in psychometric models and educational measurement", Journal of Educational Measurement, 26(1), 81-90.

Ball, W. W. Rouse. (1919), A Short Account of the History of Mathematics. New York: Macmillan.

Bohr, Niels. (1983), "Discussion with Einstein on epistemological problems in atomic physics", in Quantum Theory and Measurement. Ed. John A. Wheeler & W. Zurek. Princeton: Princeton University Press.

Brink, Nicholas 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.

Bunt, Lucas N. H., Jones, Phillip S. & Bedient, Jack D. (1976), The Historical Roots of Elementary Mathematics. Englewood Cliffs, NJ: Prentice Hall.

Burtt, Edwin A. (1954), The Metaphysical Foundations of Modern Science. New York: Doubleday Anchor.

Cajori, Florian. (1985), A History of Mathematics. New York: Chelsea.

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

Cherryholmes, Cleo. (1988), "Construct validity and the discourses of research", American Journal of Education, 96(3), 421-457.

Cohen, L. (1979), "Approximate expressions for parameter estimates in the Rasch model", British Journal of Mathematical and Statistical Psychology, 32, 113-120.

Cook, Thomas D. & Campbell, Donald T. (1979), Quasi-Experimentation: Design & Analysis Issues for Field Settings. Boston: Houghton Mifflin.

Cronbach, Lee J. (1982), "Prudent aspirations for social inquiry", in The Social Sciences: Their Nature and Uses. William H. Kruskal, ed. Chicago: University of Chicago Press.

Cronbach, Lee & Meehl, Paul. (1955), "Construct Validity in Psychological Tests", Psychological Bulletin, 52(4), 281-302.

Crouse, James & Trusheim, Dale. (1988), The Case Against the SAT. Chicago: University of Chicago Press.

Divgi, D. R. (1986), "Does the Rasch model really work for multiple choice items? Not if you look closely", Journal of Educational Measurement, 23(4), 283-296.

_____. (1989), "Reply to Andrich and Henning", Journal of Educational Measurement, 26(3), 295-299.

Douglas, Graham & Benjamin D. Wright. (1986), "The two category model for objective measurement", Research Memorandum No. 34, MESA Psychometric Laboratory, Dept. of Education, University of Chicago.

Duncan, Otis Dudley. (1984), Notes on Social Measurement: Historical and Critical. New York: Russell Sage Foundation.

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

_____. (1989), "Historical views of the concept of invariance and measurement theory in the behavioral sciences", Paper presented at the Fifth International Objective Measurement Workshop, University of California, Berkeley, March.

Fahnestock, Jeanne. (1986), "Accommodating science: The rhetorical life of scientific facts", Written Communication, 3(3), 275-296.

Fisher, William P. (1988a), "Recent developments in the philosophy of science pertaining to problems of objectivity in measurement", Rasch Measurement Transactions, 2(2), 1-3.

_____. (1988b), Truth, Method and Measurement: The Hermeneutic of Instrumentation and the Rasch Model. Unpublished dissertation, University of Chicago Department of Education.

_____. (1989), "The interdependence of educational research and practice: Cultural transmission, research methodology, and the practice of education", Paper presented at the American Educational Research Association Annual Meeting, San Francisco, March.

_____. (1990), "Conversing, testing, questioning", Paper presented at the 1990 American Educational Research Association Annual Meeting, Boston, April.

_____. (1991), "Objectivity in measurement: A philosophical history of Rasch's separability theorem", in *Objectivity in Measurement: Theory into Practice*. Mark Wilson, Ed. Norwood, NJ: Ablex.

Fleck, Ludwig. (1979), The Birth and Genesis of a Scientific Fact. introd. Thomas Kuhn. Chicago: University of Chicago Press.

Forster, Fred. (1987), Unpublished letter of rebuttal to Divgi (1986) submitted to the editor of the Journal of Educational Measurement. Gresham, Oregon

Gadamer, Hans-Georg. (1980), Dialogue and Dialectic: Eight Hermeneutical Studies on Plato. Trans. and intro. P. Christopher Smith. New Haven: Yale University Press.

_____. (1989), Truth and Method. 2d Ed. Translation revised by Joel Weinsheimer & Donald G. Marshall. New York: Crossroad.

Goldman, Steven H. & Raju, Nambury S. (1986), "Recovery of one- and two-parameter logistic item parameters: An empirical study", Educational and Psychological Measurement, 46, 11-21.

Goldstein, Harvey. (1977), "Monitoring educational standards -- An inappropriate model", Bulletin of the British Psychological Society, 30, 309-311.

_____. (1979), "Consequences of using the Rasch model for educational assessment", British Educational Research Journal, 5(2), 211-220.

_____. (1980), "Dimensionality, bias, independence and measurement scale problems in latent trait test score models", British Journal of Mathematical and Statistical Psychology, 33, 234-246.

_____. (1983), "Measuring changes in educational attainment over time: Problems and possibilities", Journal of Educational Measurement, 20(4), 369-377.

Gould, Stephen J. (1981), The Mismeasure of Man. New York: W. W. Norton.

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

Guttman, Louis. (1950), "The basis for scalogram analysis", In Measurement and Prediction. Ed. S. A. Stouffer et al. New York: John Wiley & Sons.

Hacking, I. (1983), Representing and Intervening: Introductory Topics in the Philosophy of Natural Science. Cambridge: Cambridge University Press.

_____. (1988), "On the stability of the laboratory sciences", The Journal of Philosophy, 85(10), 507-514.

Hambleton, R. K., Ed. (1983), Applications of Item Response Theory. Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R. K. & Cook, L. L. (1977), "Latent trait models and their use in the analysis of educational test data", Journal of Educational Measurement, 14(2), 75-96.

Hambleton, R. K. & Novick, M. R. (1973), "Toward an integration of theory and method for criterion-referenced tests", Journal of Educational Measurement, 10, 159-170.

Hambleton, R. K. & Rogers, H. J. (1989), "Solving criterion-referenced measurement problems with item response models", International Journal of Educational Research, 13(2), 145-160.

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

_____. (1983a), "Natural science as a hermeneutic of instrumentation", Philosophy of Science, 50(June), 181-204.

_____. (1983b), "Perception as a hermeneutical act", Review of Metaphysics, 37(1), 61-75.

_____. (1983c), Space-Perception and the Philosophy of Science. Berkeley: University of California Press.

_____. (1985), "Interpretation in physics: Observation and measurement", Greater Philadelphia Philosophy Consortium, March.

_____. (1988), "Experiment and theory: Constitution and reality", The Journal of Philosophy, 85(10), 515-524.

_____. (1989a), "After experiment: Realism and research", American Philosophical Quarterly, 26(4), 297-308.

_____. (1989b), "Hermeneutics and natural sciences: Problems and prospects--Commentary", Paper read at the American Philosophical Association (Central) Meeting, Chicago, April.

Heidegger, Martin. (1962), Being and Time. Trans. John Macquarrie and Edward Robinson. New York: Harper & Row.

_____. (1967), What is a Thing? Trans. W. B. Barton, Jr. & Vera Deutsch. Analytic afterword by Eugene Gendlin. South Bend, IN: Regnery.

Henning, Grant. (1989), "Does the Rasch model really work for multiple-choice items? Take another look: A response to Divgi", Journal of Educational Measurement, 26(1), 91-97.

Hesse, Mary. (1970), Models and Analogies in Science. Notre Dame: University of Notre Dame Press.

_____. (1972), "In defence of objectivity", Proceedings of the British Academy, 58, 275-292.

Holton, Gerald. (1988), Thematic Origins of Scientific Thought. Revised ed. Cambridge: Harvard University Press.

Hudson, L. (1972), The Cult of the Fact. New York: Harper & Row.

Ihde, Don. (1979), Technics and Praxis. Boston: D. Reidel.

Ingebo, George. (1987). Unpublished letter of rebuttal to Divgi (1986) submitted to the editor of the Journal of Educational Measurement. Gresham, Oregon

Jaeger, Richard M. (1987), "Two decades of revolution in educational measurement!?", Educational Measurement: Issues and Practice 6(2), 6-14.

Jansen, Paul G. W. (1984), "Relationships between the Thurstone, Coombs, and Rasch approaches to item scaling", Applied Psychological Measurement, 8(4), 373-383.

Karr, Chadwick. (1987), "The Rasch model does work when used properly", Unpublished rebuttal to Divgi (1986) submitted to the editor of the Journal of Educational Measurement. Portland State University, Portland, Oregon.

Kuhn, Thomas S. (1961), "The function of measurement in modern physical science", Isis 52(168), 161-193.

_____. (1970), The Structure of Scientific Revolutions. 2d ed. Chicago: University of Chicago Press.

Latour, Bruno & Woolgar, Steve. (1979), Laboratory Life: The Social Construction of Scientific Facts. Beverly Hills: Sage.

Linacre, John M. & Wright, Benjamin D. (1987), "Item bias: Mantel-Haenszel and the Rasch model", Memorandum No. 39, MESA Psychometric Laboratory, Department of Education, University of Chicago.

_____. (1989), "The equivalence of Rasch PROX and Mantel-Haenszel", Rasch Measurement Transactions, 3(2), 1-3.

Lindquist, E. F. (1953), "Selecting appropriate score scales for tests (Discussion)", Proceedings of the 1952 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

Loevinger, Jane. (1957), "Objective tests as instruments of psychological theory", Psychological Reports, 3, 635-694.

_____. (1965), "Person and population as psychometric concepts", Psychological Review, 72(2), 143-155.

London, Fritz & Bauer, Edmond. (1983), "The theory of observation in quantum mechanics", in Quantum Theory and Measurement. Ed. John A. Wheeler & W. Zurek. Princeton: Princeton University Press. Reprinted from "La theorie de l'observation en mecanique quantique", No. 775 of Actualites Scientifiques et industrielles: Exposes de Physique Generale. Publies sous la direction de Paul Langevin. Hermann, Paris, 1939.

Lord, Frederic M. (1953), "On the statistical treatment of football numbers", The American Psychologist, 8, 750-751.

_____. (1968), "An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model", Educational and Psychological Measurement, 28, 989-1020.

_____. (1975), "Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters", Research Bulletin 75-33. Princeton, New Jersey: Educational Testing Service.

_____. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, New Jersey: Erlbaum.

_____. (1983), "Small N justifies Rasch model", in New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing. Edited by David J. Weiss. New York: Academic.

Lord, Frederic M. & Novick, M. R. (1968), Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley.

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

Messick, Samuel. (1975), "The standard problem: Meaning and values in measurement and evaluation", American Psychologist 30(October), 955-966.

Mislevy, R. J. & Bock, R. D. (1983), BILOG: Item Analysis and Test Scoring with Binary Logistic Models. Mooresville, Indiana: Scientific Software, Inc.

Ormiston, Gayle & Sassower, Raphael. (1989), Narrative Experiments: The Discursive Authority of Science and Technology. Minneapolis: University of Minnesota Press.

Osburn, H. G. (1968), "Item sampling for achievement testing", Educational and Psychological Measurement, 28, 95-104.

Owen, David S. (1985), None of the Above: Behind the Myth of Scholastic Aptitude. Boston: Houghton Mifflin.

Petersen, Aage. (1968), Quantum Physics and the Philosophical Tradition. Cambridge: MIT Press.

Phillips, S. E. (1986), "The effects of the deletion of misfitting persons on vertical equating via the Rasch model", Journal of Educational Measurement, 23(2), 107-118.

Phillips, S. E. & Mehrens, W. A. (1987), "Curricular differences and unidimensionality of achievement test data: An exploratory analysis", Journal of Educational Measurement, 24(1), 1-16.

Polanyi, Michael. (1958), Personal Knowledge: Towards a Post-Critical Philosophy. Chicago: University of Chicago Press.

_____. (1967), The Tacit Dimension. Garden City, NJ: Doubleday.

Rasch, Georg. (1960), Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danmarks Paedogogiske Institut; reprint, 1980, with Foreword and Afterword by Benjamin D. Wright, Chicago: University of Chicago Press.

_____. (1961), "On general laws and the meaning of measurement in psychology", Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 4, 321-333 Berkeley: University of California Press.

Rentz, R. R. & W. L. Bashaw. (1977), "The national referen[illegible]e for reading: An application of the Rasch model", Journal of Educational Measurement, 14, [illegible].

Ricoeur, Paul. (1965), History and Truth. Trans. Charles A. Kelbley. Evanston: Northwestern University Press.

_____. (1981), Hermeneutics and the Human Sciences: Essays on Language, Action and Interpretation. Ed., trans. and intro. by John B. Thompson, with a response from the author. Cambridge: Cambridge University Press.

Rorty, Richard. (1985), "Solidarity or Objectivity", in Post-Analytic Philosophy, edited by John Rajchman & Cornell West. New York: Columbia University Press.

Singleton, Margles. (1991), "Rasch measurement as a Kuhnian revolution", Rasch Measurement Transactions, 4(4), 119.

Smith, Richard. (1990), "Recovering pseudo-item banks", Paper presented to the Autumn Midwest Objective Measurement Seminar, University of Chicago.

Stenner, A. Jackson & Smith, Malbert III. (1982), "Testing construct theories", Perceptual and Motor Skills, 55, 415-426.

Stenner, A. Jackson, Smith, Malbert III, & Burdick, Donald S. (1983), "Toward a theory of construct definition", Journal of Educational Measurement, 20(4), 305-316.

Stevens, S. S. (1946), "On the theory of scales of measurement", Science, 103, 677-680.

Stocking, Martha L. (1989), "Empirical estimation errors in item response theory as a function of test properties", Educational Testing Service Research Report, Princeton, New Jersey.

Strenio, Andrew J. (1981), The Testing Trap. New York: Rawson, Wade.

Sutherland, Gillian, in collaboration with Stephen Sharp. (1984), Ability, Merit, and Measurement: Mental Testing and English Education, 1880-1940. Oxford: Clarendon Press.

Thurstone, L. L. (1928), "Attitudes can be measured", American Journal of Sociology, 33, 529-554. Reprinted in L. L. Thurstone. The Measurement of Values. Chicago: University of Chicago Press, Midway Reprint Series, 1959.

Toulmin, Stephen. (1953), The Philosophy of Science: An Introduction. New York: Harper & Row.

_____. (1982), "The construal of reality: Criticism in modern and postmodern science", Critical Inquiry 9(September), 93-111.

Tracy, David. (1975), Blessed Rage for Order: The New Pluralism in Theology. Minneapolis: The Winston-Seabury Press.

Vygotsky, Lev. 1978. Mind in Society: The Development of Higher Psychological Processes. Cambridge: Harvard University Press.

Wheeler, John A. (1983), Law without law. In Quantum Theory and Measurement. Ed. John A. Wheeler & W. Zurek. Princeton: Princeton University Press.

Whitely, Susan E. & Rene V. Dawis. (1974), "The nature of objectivity with the Rasch model", Journal of Educational Measurement, 11(2), 163-178.

Whitely, Susan E. (1977), "Models, meanings and misunderstandings: Some issues in applying Rasch's theory", Journal of Educational Measurement, 14(3), 227-235.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982), LOGIST User's Guide. Princeton, New Jersey: Educational Testing Service.

Wood, Robert. (1978), "Fitting the Rasch model: A heady tale", British Journal of Mathematical and Statistical Psychology, 31, 27-32.

Wright, Benjamin D. (1968), "Sample-free test calibration and person measurement", in Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton: Educational Testing Service, pp. 85-101.

_____. (1977a), "Misunderstanding the Rasch model", Journal of Educational Measurement, 14(3), 219-225.

_____. (1977b), "Solving measurement problems with the Rasch model", Journal of Educational Measurement, 14(2), 97-116.

_____. (1984), "Despair and hope for educational measurement", Contemporary Education Review, 3(1), 281-288.

_____. (1985), "Additivity in psychological measurement", in Measurement and Personality Assessment. Ed. Edward Roskam. North Holland: Elsevier Science Publishers.

_____. (1988), "The model necessary for a Thurstone scale" and "Campbell concatenation for mental testing", Rasch Measurement Transactions, 2(1), 2-4.

_____. (1989), "Deducing the Rasch model", Unnumbered MESA Research Memorandum, MESA Psychometric Laboratory, Dept. of Education, University of Chicago.

Wright, Benjamin D. & Bell, Susan R. (1984), "Item banks: What, why, how", Journal of Educational Measurement, 21(4), 331-345.

Wright, Benjamin D. & Douglas, Graham. (1986), "The rating scale model for objective measurement", Research Memorandum No. 35, MESA Psychometric Laboratory, Dept. of Education, University of Chicago.

Wright, Benjamin D., Linacre, John M., & Schultz, Matthew. (1990), BIGSCALE: A Rasch-Model Rating Scale Analysis Computer Program. Chicago: MESA Press.

Wright, Benjamin D. & Masters, Geofferey. (1982), Rating Scale Analysis. Chicago: MESA Press.

Wright, Benjamin D. & Stone, Mark. (1979), Best Test Design. Chicago: MESA Press.