DOCUMENT RESUME

ED 295 961 TM 011 738

AUTHOR Engelhard, George, Jr.
TITLE Thorndike's and Wood's Principles of Educational Measurement: A View from the 1980's.
SPONS AGENCY National Academy of Education, Washington, D.C.
PUB DATE Apr 88
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational History; *Educational Testing; Testing Problems; *Test Theory

ABSTRACT

The purpose of this essay is to describe the principles of educational measurement proposed by B. Wood during the 1920s in his dissertation, written under the direction of E. L. Thorndike, and later published as "Measurement in Higher Education" (1923) These principles were selected because they illustrate one of the earliest and most complete descriptions of a set of basic and perennial problems encountered in educational testing. The specific questions addressed in this essay are concerned with the following: (1) the basic measurement problems identified by Thorndike and Wood in the first two decades of this century; (2) the means by which these measurement problems appear within the context of educational testing according to Wood; (3) means by which these problems were addressed by Wood in the 1920s; and (4) contemporary views of these problems. Principles of educational measurement (objectivity, defined zero and unit, definition of the function to be measured, consistency, within person variability, comparability, distinctness of power and achievement, equal exposure and practice, advantages of indirect measurement, test construction, test use, and measurement must not be confused with pedagogy) are tabulated according to specific problems and proposed solutions to each. Nine pages of references are provided. (Author/THJ)

THORNDIKE'S AND WOOD'S PRINCIPLES OF EDUCATIONAL MEASUREMENT: A VIEW FROM THE 1980'S

George Engelhard, Jr.

Emory University

Address: Professor George Engelhard, Jr.
Emory University
Division of Educational Studies
210 Fishburne Building
Atlanta, GA 30322

(404) 727-0607 (work)
(404) 982-0616 (home)

Running head: PRINCIPLES OF EDUCATIONAL MEASUREMENT

Paper presented at the annual meeting of the American Educational Research Association in New Orleans, April 1988

Abstract

The purpose of this essay is to describe the principles of educational measurement proposed by Ben Wood during the 1920's in his dissertation which was written under the direction of E. L. Thorndike, and later published as Measurement in Higher Education (1923). These principles were selected because they illustrate one of the earliest and most complete descriptions of a set of basic and perennial problems encountered in educational testing. The specific questions addressed in this essay are as follows: What were the basic measurement problems identified by Thorndike and Wood in the first two decades of this century? How do these measurement problems appear within the context of educational testing according to Wood? How were these problems addressed by Wood in the 1920's? And how are these problems viewed today?

THORNDIKE'S AND WOOD'S PRINCIPLES OF EDUCATIONAL MEASUREMENT: TEST THEORY IN THE 1920'S[1]

> The history of science is the history of measurement. (Cattell, 1893, p.316)

> What ever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality. (Thorndike, 1918, p.16)

In his presidential address to NCME last year, Jaeger (1987) reminded the educational measurement community of the importance of periodically reviewing the history of our discipline. He eloquently summed up his remarks as follows:

> I would assert that to move forward efficiently we must first look back -- to incorporate and build upon the riches of the past while avoiding futile paths earlier explored and appropriately abandoned. To dwell on the past is folly; to ignore it is absurdity. (p. 13)

This essay is intended to identify what I consider one source of these "riches". Specifically, I would like to discuss a fairly complete theory of educational testing proposed by Ben Wood in 1923 based on the measurement theory of E. L. Thorndike (1904; 1919). These principles were selected because they can be used to illustrate some of the basic and perennial problems encountered in

educational measurement, and also provide a useful framework for the exploration of its history.

What were the basic measurement problems identified by Thorndike and Wood in the first two decades of this century? How do these measurement problems appear within the context of educational testing according to Wood? And how were these problems addressed by Wood in the 1920's? This essay is intended to provide answers to these questions. Some brief comments will also be made on how these principles appear today.

Thorndike and Wood

In 1904, E. L. Thorndike published the first edition of his highly influential book entitled An Introduction to the Theory of Mental and Social Measurements. Thorndike's major aim in writing the book was to

> . . . introduce students to the theory of mental measurements and to provide them with such knowledge and practice as may assist them to follow critically quantitative evidence and argument and to make their own researches exact and logical
>
> (Thorndike, 1919, p. v)

Thorndike's book was the standard reference on statistics and quantitative methods in the mental and social sciences for the first two decades of this century (Clifford, 1984; Travers, 1983). Much of this influence can be attributed to Thorndike's clear and

expository writing style. He explicitly acknowledges that the then current work in measurement theory had not been presented in a manner suitable for students without fairly advanced mathematical skills, and he set out to present a less mathematical introduction to measurement theory based on the belief that "there is, happily, nothing in the general principles of modern statistical theory but refined common sense, and little in the techniques resulting from them that general intelligence can not readily master" (p. 2). Many of us that have struggled with the mathematics of item response theory can appreciate Thorndike's comments, and applaud his attempt.

Although Thorndike wrote extensively on educational measurement, covering topics which ranged from the general statement of his theory (Thorndike, 1904; 1919) to the measurement of a variety of educational outcomes (Thorndike, 1910, 1914, 1921), as well as intelligence (Thorndike, et al., 1926), I have found that one of the clearest and most complete statements of Thorndike's measurement theory was presented by his student and colleague, Ben Wood. In a chapter titled "Some Principles of Educational Measurement", Wood (1923) stated that

> This chapter is little more than an effort to expand that treatment [of measurement theory] for the purpose of exposition. Practically all the material in this chapter

> is taken from Professor Thorndike's well-known treatise, or directly inferred from some of its propositions.
>
> (Wood, 1923, p. 141)

Wood has provided a careful and useful exegesis of Thorndike's early work on measurement and its implications for educational testing. Wood's work provides the structure for discussing the principles of educational measurement presented here. 2

What were the basic measurement problems identified by Thorndike and Wood? Thorndike clearly stated that the "special difficulties" of measurement in the behavioral sciences are

(1) Absence or imperfection of units in which to measure;

(2) Lack of constancy in the facts measured;

(3) Extreme complexity of the measurements to be made.

In order to illustrate the problems related to the absence of an accepted unit or measurement, Thorndike (1919) pointed out that the spelling tests developed by Joseph Mayer Rice (Graham, 1966) did not have equal units. Rice assumed that all of his spelling words were of equal difficulty, while Thorndike argued that the correct spelling of an easy versus a hard word did not reflect equal amounts of spelling ability. Because the units of measurement are unequal, Thorndike asserted that Rice's results were inaccurate. Without general agreement on units, the meaning of our test scores become more subjective.

Inconstancy is the second major measurement problem identified by Thorndike (1919). Many of the measurement problems encountered in the behavioral sciences are related to the random variation inherent in many human characteristics. Not only are these variations due to the unreliability of our tests, but they also reflect within subject fluctuations. For example, if we measure a person's motivation, or even body temperature repeatedly, these values tend to vary.

The final measurement problem or "special difficulty" identified by Thorndike pertains to the extreme complexity of the variables and constructs that we wish to measure. Most of the variables worth measuring in the behavioral sciences, such as mathematics ability, intelligence, competitiveness, do not readily translate into unidimensional tests which permit the reporting of a single score to represent the individual's location on the variable.

Some Principles of Educational Measurement

In addressing the three "special difficulties" identified by Thorndike within the context of education, Wood (1923) identified a set of sixteen principles which included technical recommendations on test construction, as well as more policy-oriented issues related to test use in education. One of Wood's major concerns was

with how the new objective items could be used to solve a number of measurement problems in higher education. A summary of Wood's principles, problems and proposed solutions is given in Table 1.

Insert Table 1 about here

I should also point out, as Wood did, that these principles were intended to be taken in concert as solutions to the three problems in measurement identified by Thorndike. In the following sections, each of Wood's principles will be presented and discussed.

Objectivity

Both Thorndike and Wood considered objectivity to be one of the most important characteristics of a valid test. According to Thorndike (1919), "a perfectly objective scale is a scale in respect to whose meaning all competent thinkers agree" (p. 141). How can agreement on the meaning of the scores on a test be obtained? Thorndike (1919) proposed the creation of a set of standard items calibrated onto a scale which would be used as a "common measuring stick", while Wood (1923) addressed this measurement problem in terms of the objectivity of the scoring method. To quote Wood (1923), "the True-False test is a good example of an objective mental scale. No competent person would

disagree in rating a True-False paper, provided they used the key which accompanies the test" (p. 144). Anticipating the idea that reliability is necessary but not sufficient to establish the validity of a test, Wood (1923) stated that "it is perfectly possible to have a very objective scale without having one which measures the facts to be measured" (p. 144).

From a current perspective, Wood (1923) clearly was dealing with a problem related to the reliability of the test scores, although the more general view of this principle based on Thorndike (1919) suggests that Thorndike also included aspects associated with validity. The meaning of test scores, and any consensus about their meaning, would involve establishing both their reliability and validity. Many current measurement textbooks use the term "objectivity" of scoring much as Wood did (Anastasi, 1988; Cronbach, 1984). Further, the word "objectivity" is used in another way in the measurement theory of Georg Rasch (1977, 1980). According to Wright and Stone (1979), two conditions are necessary for objectivity as viewed by Rasch, and these are (1) the calibration of the measurement must be independent of those objects that happen to be used for the calibration and (2) the measurement of objects must be independent of the instrument that happens to be used for measuring.

Reference to a defined zero point

in terms of a defined unit

One of the key problems in educational measurement is the establishment of scaling methods which provide meaningful and interpretable test scores. The solution to this problem according to Thorndike and Wood was based on the development of scales with defined zero points, either arbitrary or absolute, and the selection of a stable unit of measurement. The solution proposed by Wood (1923) was based on a z-score transformation with the mean defining an arbitrary zero point, and the standard deviation as the unit. Wood selected the mean and standard deviation because of their relative "stability". Wood (1923), also dealt with another aspect of the scaling problem related to the comparability of test scores which would be viewed today as an equating problem. In his words,

> the same test applied to different groups gives both different points of origin and different Standard Deviations. Universal comparisons can therefore be made only when measurements are expressed in terms of the Standard Deviation (and reckoned from the Mean), of some defined and standard distribution. (Wood, 1923, p. 150).

Current approaches to the problem of scaling include a whole array of methods for equating based on classical test theory and item

response theory (Brennan, 1987; Skaggs & Lissitz, 1986; Yen, 1986).

The Principle of Definition

In his third principle, Wood (1923) returned to a question connected to the validity of the test scores. Validity refers to the appropriateness, meaningfulness and usefulness of the inferences which can be made from the test scores (Standards for Educational and Psychological Testing, 1985). The basic question is as follows: What is the test actually measuring? Wood (1923) proposed that a precise operational definition of the construct be used to answer this question, and that this definition would make clear what the test measured. Wood's view here is close to the modern idea of content validity which is not too surprising given his focus on educational achievement tests. Neither Thorndike nor Wood, included the broader validity issues implied by the question raised in this section — what is the test measuring? — which would include obtaining criterion-related and construct-related evidence relevant to this question. Recent arguments have been made for the importance of construct validity as well as content validity for achievement tests (Haertel, 1985). Under the principle of definition, Wood (1923) also anticipated problems related to the development of operational definitions for complex [illegible] as reading achievement and intelligence.

Consistency

A recurring problem in educational and psychological measurement relates to the complexity of the constructs that we wish to measure. This "extreme complexity" is dealt with by Wood (1923) in terms of the concept of "consistency" which might be called unidimensionality today. In Wood's example, he points out that a "notable example of obvious impurity of measurement is afforded by some arithmetic tests . . . problems in these tests are but little more than very severe reading tests . . . it would seem more advantageous for all purposes of measurement to separate the two functions" (pp. 154-155). Unfortunately, the dimensionality of a set of test items can not be adequately assessed simply by examining the content of the items. How do we really know that when an individual responds to a set of test items, he or she is really only using one ability or many? In many instances, useful test scores are produced by summing what Thorndike and Wood might view as "inconsistent items". The early Binet and Simon test was criticized on this basis by Spearman (1927), who referred to their intelligence test as a set of "hotchpot procedures" (p. 66). Wood did not have adequate procedures for dealing with this problem, and exciting current work in item factor analysis (Bock, Gibbons and Muraki, in press; Mislevy, 1986; Muthen, 1984) s contributed to the problem of assessing the "consistency" or dimensionality of our

tests in education and psychology.

Within Person Variability

In this fifth principle, Wood (1923) is concerned with the problem of "variability of mental functions in the same individual from day to day and hour to hour" (p.155). Usually, we think of individuals as having fairly stable behavioral characteristics. These characteristics are not really fixed, but can be viewed as an average over a number of observations. This intra-individual variability may be due to a variety of factors, such as boredom, anxiety, fatigue or illness, and must be taken into account in measurement. If the intra-individual variability in responses is great, then the problem of identifying differences between individuals becomes more difficult. In order to address this problem, Wood (1923) recommends administering as many items as possible. In his words, "Only by taking a large sample of an individual's performances can we arrive at a reliable estimate of his normal or average ability" (p. 151). When a more complex variable is measured, such as reading ability, then it is even more important to increase the number of items. Wood (1923) referred to this issue as the "principle of increasing accuracy" (p. 151). This principle would be viewed from a current perspective as dealing with the reliability of the scores and the standard error of measurement which provides an index of this response

variability. It is well known, all else being equal, that we can increase the reliability of test scores by increasing the number of items because a better sample of the content domain can be obtained. Generalizability theory (Cronbach, et al., 1972) provides an approach which can be used to examine various sources of random variation which can be useful in addressing this measurement problem.

Comparability

This principle deals with a problem related to test use. The word "comparability" is used because Thorndike and Wood believed that once a test had been calibrated, the application of this test involved a comparison between the test and the person to be measured. This idea can be visualized more clearly if we think of the problem of measuring writing ability using a standard set of essays. Once these essays have been calibrated from poor to excellent, a judge "compares" each new essay to the set of standards in order to define the level of writing ability reflected in each essay. This measurement problem relates to the question of whether or not the test can be validly applied with reasonable ease and accuracy to the objects being measured. As an example, a bathroom scale is not accurate enough to use in weighing gold. Wood's proposed solution was to select an appropriate test to measure the construct of interest. In grading an essay, the topics

addressed in the calibrated essays should be the same as those the examinees are writing about.

Power and Achievement

Wood stressed that the distinction between "power" (intelligence) and "achievement" must be kept in mind in the construction, administration, and interpretation of all test results. In order to illustrate this principle, he uses as an example the problem of placing two students with very different backgrounds, one from a rural setting and the other from an urban setting, in reading ability groups. The reading achievement score of the urban child was higher than the rural child's score, and the teacher planned to place the rural child in the lowest reading group. Additional information was available on the Terman intelligence test which indicated that the rural child had an I.Q. of 130 and at the urging of Wood, she was placed in a higher reading group. The subsequent reading achievement was quite high. The major point here seems to be that these two types of test can provide different information about an individual differences, and that this information can be useful in educational planning and decision making. Recent views of intelligence testing suggests that the distinction between intelligence and achievement as measured by current IQ tests may not be as clear as previously believed (Anastasi, 1983). Further, many intelligence test used

today in schools have been renamed as tests of "school ability", "scholastic ability" and "academic aptitude" (Beck, 1986) which indicates that these types of tests, whatever they are called, can serve important functions for education similar to those envisioned by Wood.

Principle of Equal Exposure and Practice

How can differences in opportunity to learn be addressed when testing general intelligence? In answering this question Wood (1923) stated that "inferences as to the general intelligence or inborn ability of two individuals must be based upon their reactions to material to which they have been equally exposed and in which they have had equal practice, except insofar as exposure and practice are influenced by native capacity" (p. 158). In order to minimize the effects of opportunity to learn, Wood recommended that "emphasis should be placed upon testing mental processes which are largely independent of informational content" (p. 160), while recognizing that differences in past exposure can never by completely eliminated. In situations where there are large differences in the home or social environment, these must be considered in explaining differences in achievement, general intellect and special abilities. Wood's views are fairly modern, although he does seem to be a bit optimistic about the possibility of controlling for these differences in opportunity to learn. The

problem of exposure and practice is still an important issue because it can have a significant impact on the way in which test scores are interpreted. If two children are tested on a common set of educational objectives, and one has not had the opportunity to learn the objectives is it fair to compare children on this test? Do the scores have the same meaning? This problem is reflected in current issues related to customized testing (Yen, Green & Burket, 1987), and curricular validity (Mehrens & Phillips, 1987). There seems to be general agreement that if opportunity is an important factor, then it must be taken into account in the interpretation of the test scores, however, the methods for doing this are still the subject of debate.

Advantages of Indirect Measurement

This principle treats two related problems -- What are the advantages of objective items which Wood (1923) called "indirect measurement" as compared to essay items? Or more broadly conceived, what is the best type of item to measure a construct? This principle is concerned with the disadvantages of essays as an item type, and Wood's advocacy of "new-type or objective" examinations in education. According to Wood (1923), "the essay examination in the hands of the average teacher does measure a very important element which apparently cannot be measured directly by any other means thus far developed. But it measures that element

very incompletely and very unreliably" (p. 161). Wood (1923) identified two major weaknesses in essay exams, (1) inadequate sampling of examinee performance and of material and (2) variability and uncertainty in the subjective methods used to score essays. He presents a strong case for the use of objective items types, such as completion, true-false and recognition items, in higher education, and recommends that "where indirect methods have demonstrable advantages over direct measurements, indirect measurements should be used" (p. 151). In spite of his arguments against essay type items, he still felt that essay items played an important role. In his words, "indirect measurement is not suggested as a substitute for, but as a supplement to, direct measurement" (p. 161). From the perspective of the 80's, both methods would be viewed as "indirect" as opposed to "performance-type" tests (Anastasi, 1988). There is little if any debate over the usefulness of "indirect measures", such as multiple choice items today. The debate today centers on when a particular item type is appropriate. Although Wood (1923) was discussing the use of essays to assess achievement in the content areas, a similar set of concerns appear today in the use of essays to measure writing ability. Essay type items and the assessment of writing ability are being increasingly used in state and national assessment programs as well as a part of standardized achievement tests

(Quellmalz, 1986; Special Issue on Writing Assessment, 1984), although scoring issues remain (Chase, 1986).

Test Construction

Wood's next 3 principles are related to test construction, and the steps in test construction that can increase the validity of the test scores. According to Wood, (1) "a valid test must contain a larger number of small elements" (p. 163), (2) "in the measurement of any mental function as many types of questions should be employed as administrative conditions allow" (p. 165) and (3) "the questions should involve as little as possible irrelevant considerations and superfluous activities on the part of the examinee" (p. 168). The principle of "many small elements" in (1) above reflects Wood's case against the use of essay items. Most educational tests created today do not follow the recommendation regarding the use of multiple item types as suggested in (2) above. Gulliksen (1986) has attributed this to "the failure to distinguish between the requirements of standardized testing and classroom testing seems to be responsible for the lack of improvement--and perhaps even a decline—in the quality of teacher-made classroom tests over the past 40 years" (p. 189). Gulliksen (1986) goes on to call for the use of a variety of item types by teachers. Wood (1923) recommended seven conditions for constructing a "good" item which are commonly recommended today within standard texts on

educational testing. For example, Wood suggested that the items should not contain "trick" elements and chance influences should be minimized. A current topic in test construction not directly addressed by Thorndike and Wood is how to detect item bias (Linn & Drasgow, 1987; Shepard, Camilli & Williams, 1985). Given the social context of their work, this was simply not a measurement issue at the time. Cronbach (1975) and Haney (1981) provide interesting and useful discussions of the interplay between social concern, policy and testing.

Test Use

The next 3 principles refer generally to test use, and the match between persons and items in terms of appropriateness. Wood was concerned with the adequacy of a test in terms of measuring the whole range of a construct for a particular group. The problem would be evident if the test was too easy or too hard for the examinees, and the test scores would not be distributed on the variable. In other words, the test would not be able to detect individual differences -- the sine qua non of measurement. If the test is "appropriate" then "it must be sensitive to and capable of registering real differences in every part of the range of the quality it is designed to measure" (p. 171). Further, Wood (1923) points out that "no absolute criterion is available to show whether an exam fully satisfies this condition, but fairly secure indices

are not wanting" (p. 170). Wood also had a parallel concern with the distribution of the items on the scale. The measurement goal is to develop a set of items that are in the appropriate range of difficulty for a group of examinees. When an individual encounters an "inappropriate" item, guessing and other chance influences can interfere with the measurement process. Wood pointed out that "chance influences must be recognized and countered in the construction, scoring, and evaluation of every type of question" (Wood, 1923, p. 172). The concerns expressed by Wood could not be handled adequately with the methods available in the 1920's. From a current perspective, Wood's concerns here could be examined with a "map of the variable" (Wright and Stone, 1979) which provides a graphic display which shows simultaneously the location and distribution of items and individuals on the variable. Further, recent work on computerized adaptive testing (Green, et al., 1984; Weiss, 1982) is explicitly motivated by this concern with the match between items and individuals in terms of appropriate item difficulty. Additional work on appropriateness indices provides another approach which can be used to examine the validity of individual test scores (Drasgow, Levine, & Williams, in press).

Measurement and Pedagogy

Wood (1923) believed that in the construction and administration of examinations, measurement must not be confused

with pedagogy. Wood is again defending the use of objective items which had been apparently criticized as having no pedagogical value as compared to essay items. His major point is that although both types of items can have pedagogical value, the value of an examination must be assessed separately in regards to these two issues. According to Wood (1923), "intrinsic pedagogical value in an examination is highly desirable, but the value of the examination as a measuring device cannot be made to depend on its value as a teaching device (p. 174). Today many uses of tests involve the explicit development of a link between testing and instruction (Airasian & Madaus, 1983; Burstein, 1983; Glaser, 1986; Stiggins, Conklin & Bridgeford, 1986).

Discussion and Implications

In many ways, we have made a great deal of progress in psychometrics that Thorndike and Wood could not have anticipated. Recent advances in measurement theory (item response theory, generalizability theory and factor analysis), computer technology (computerized adaptive testing, video discs), and statistical methodology (probabilistic models for analyzing qualitative data and Bayesian methods) make possible solutions to many of our measurement problems which were undreamed of in the 1920's. And yet considering the basic measurement problems identified by Thorndike and the principles of educational measurement proposed by

Wood, it is hard not to be impressed, and it is easy at times to forget, that many of these ideas were first expressed by Thorndike almost 85 years ago and by Wood over 65 years ago.

The "special difficulties" of measurement in the behavioral sciences are still present today. Generally agreed upon units are not available for many variables of interest, human characteristics still show random variation, and there is little doubt that the variables which we wish to measure are still complex. What seems to have changed the most is not the basic questions or problems of measurement, but our ingenuity and technical finesse in finding new solutions for old problems. Although in some cases, early solutions used by Thorndike, such as item scaling, worked remarkably well (Engelhard, 1984). Classical test theory was still in its infancy when Thorndike and Wood conducted their research and proposed their measurement theories, and modern measurement theories, such as item response theory and generalizability theory were of course not developed yet. Many of the basic problems in measurement were identified at the beginning of this century, while the solutions offered have changed over time as new measurement theories are created.

It is hoped that this essay will generate some additional interest in the history of educational test theory. For example Haney and Reidy (1987) report finding only seven references that

deal directly with the history of educational testing in America in the entire ERIC data base, and I have had a similar experience with PsychLIT which is based on Psychological Abstracts. Early books on this topic, such as Linden and Linden (1968) and DuBois (1970) are now about 20 years old. Several historical articles on mental testing (Clarke & Clarke, 1985), educational testing (Resnick, 1982), educational assessment (McArthur, 1987), educational evaluation (Madaus, Stufflebeam, & Scriven, 1983) and employment testing (Hale, 1982) are available, but no book-length treatment has been published recently. Sokal (1987) has edited a volume on psychological testing and American society, and has made some concrete suggestions about approaches to the history of psychological testing (Sokal, 1984). In her recent review of two new books on the history of statistics (Porter, 1986; Stigler, 1986), Cowan (1987) has made an important distinction between histories of a discipline written by insiders versus outsiders. There is a clear need for both versions, but an updated history of the key ideas underlying measurement theory which does for psychometrics what Stigler (1986) as an "insider" has done for statistics is required. Since test theory is approaching its century mark, if we consider the Cattell article in 1893 as its birth, it would seem that a comprehensive history is somewhat overdue. I'm currently working on a project with the generous

support of a Spencer Fellowship from the National Academy of Education which focuses on the comparative and historical development of several measurement theories which I hope will contribute to this history of test theory.

In conclusion, I hope that this essay illustrates some of the insights that can be gained from a careful analysis of earlier work on educational testing. In presenting the measurement problems, I have not intended to provide an exhaustive discussion of how these problems would be addressed within the context of the major modern measurement theories, such as item response theory (Lord, 1980; Wright and Masters, 1982), generalizability theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972; Brennan, 1983) or factor analysis (Joreskog & Sorbom, 1986). Many of the problems identified by Thorndike (1919) and Wood (1923) could be the basis of articles in themselves, and my goal has been to provide a general overview, rather than a great deal of depth.

Jaeger (1987) in his presidential address posed the following question: Where's the revolution!? One partial answer is that we have not had a revolution, but maybe some "evolution", in terms of the measurement problems we seek to solve. Another answer might be that in some areas, our new theories of measurement and technological advances which deal with these problems are indeed revolutionary when viewed from the perspective of the 1920's!

References

Airasian, P. W. & Madaus, G. F. (1983). Linking testing and instruction. Journal of Educational Measurement, 20, 103-118.

Anastasi, A. (1983). What do intelligence tests measure? In S. B. Anderson & J. S. Helmich (Eds.), On educational testing. San Francisco, CA: Jossey-Bass Publishers.

Anastasi, A. (1988). Psychological testing. Sixth Edition. New York: Macmillan Publishing Co., Inc.

Beck, M. D. (1986). The Otis and Otis-Lennon tests: Their contributions. Educational Measurement: Issues and Practice, 5, 12-18.

Bock, R. D., Gibbons, R. & Muraki, E. (In press). Full-information item factor analysis. Applied Psychological Measurement.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City: ACT Publications.

Brennan, R. L. (Ed.). (1987). Problems, perspectives and practical issues in equating. [Special series]. Applied Psychological Measurement, 11.

Burstein, L. (1983). A word about this issue. Journal of Educational Measurement, 20, 99-101.

Cattell, J. K. (1893). Mental measurement. Philosophical Review, 2, 373-380.

Chase, C. I. (1986). Essay test scoring: Interaction of relevant variables. Journal of Educational Measurement, 23, 33-41.

Clarke, A. D. B. & Clarke, A. M. (1985). Mental testing: Origins, evolution and present status. History of Education, 14, 263-272.

Clifford, G.J. (1984). Edward L. Thorndike: The sane positivist. Middleton, CT: Wesleyan University Press. (Originally published 1968).

Cowan, R. S. (1987). Review of The rise of statistical thinking 1820-1900 by Porter and The History of Statistics: The measurement of uncertainty by Stigler. Journal of the American Statistical Association, 82, 1178-1179.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. American Psychologist, 30, 1-14.

Cronbach, L. J. (1984). Essentials of psychological testing. Fourth Edition. New York: Harper & Row, Publishers.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Downey, M. T. (1965). Ben D. Wood: Educational Reformer. Princeton, NJ: Educational Testing Service.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (in press). Detecting inappropriate test scores with optimal and practical indices. Applied Psychological Measurement, 11 .

DuBois, P. H. (1970). A history of psychological testing. Boston: Allyn and Bacon, Inc.

Engelhard, G. (1984). Thorndike, Thurstone and Rasch: A comparison of their methods of scaling psychological and educational measurement. Applied Psychological Measurement, 8, 21-38.

Glaser, R. (1986). The integration of instruction and testing. In The redesign of testing for the 21st century. Princeton, NJ: Educational Testing Service.

Graham, P. (1966). Joseph Mayer Rice as a founder of the progressive education movement. Journal of Educational Measurement, 3, 129-133.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L. & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive testing. Journal of Educational Measurement, 21, 347-360.

Gulliksen, H. (1986). Perspective on educational measurement. Applied Psychological Measurement, 10, 109-132.

Haertel, E. (1985). Construct validity and criterion-referenced testing. Review of Educational Research, 55, 23-46.

Hale, M. (1982). History of employment testing. In A. K. Wigdor & W. R. Garner (Eds.), Ability testing: Uses, consequences, and controversies. Part II: Documentation. Washington, DC: National Academy Press.

Haney, W. (1981). Validity, vaudeville and values: A short history of social concerns over standardized testing. American Psychologist, 36, 1021-1034.

Haney, W. & Reidy, E. F. (1987). Editorial. Educational Measurement: Issues and Practice, 6, 4-5.

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

Joreskog, K. G. & Sorbom, D. (1986). LISREL VI: Analysis of linear structural relationships by maximum likelihood, instrumental variables and least squares methods. Fourth Edition. Mooresville, IN: Scientific Software, Inc.

Linden, K. W. & Linden, J. O. (1968). Modern mental measurement: A historical perspective. Boston: Houghton Mifflin Co.

Linn, R. L. & Drasgow, F. (1987). Implications of the Golden Rule for test construction. Educational Measurement: Issues and Practice, 6, 13-17.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Madaus, G. F., Stufflebeam, D. & Scriven, M. S. (1983). Program evaluation: A historical overview. In G. F. Madaus, M. S. Scriven & D. Stufflebeam (Eds.), Evaluation models: View points on educational and human services evaluation. Boston: Kluwer-Nijhoff Publishing.

McArthur, D. L. (1987) Educational assessment: A brief history. In D. L. McArthur (Ed.), Alternative approaches to the assessment of achievement. Boston: Kluwer Academic Publishers.

Mehrens, W. A. & Phillips, S. E. (1987). Sensitivity of item difficulties to curricular validity. Journal of Educational Measurement, 24, 357-370.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.

Muthen, B. (1984). A general structural equation model for dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49, 115-132.

Porter, T. M. (1986). The rise of statistical thinking: 1820-1900. Princeton, NJ: Princeton University Press.

Quellmalz, E. S. (1986). Writing skills assessment. In R. A. Berk (Ed.), Performance assessment: Methods and applications, 492-508. Baltimore: Johns Hopkins University Press.

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

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press. [Originally published in 1960 by the Danish Institute for Educational Research].

Resnick, D. (1982). History of educational testing. In A. K. Wigdor & W. R. Garner (Eds.), Ability testing: Uses, consequences, and controversies. Part II: Documentation. Washington, DC: National Academy Press.

Shepard, L. A., Camilli, G. & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Skaggs, G. & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. Review of Educational Research, 56, 495-529.

Sokal, M. M. (1984). Approaches to the history of psychological testing. History of Education Quarterly, Fall, 419-430.

Sokal, M. M. (1987). Psychological testing and American Society. New Brunswick, NJ: Rutgers University Press.

Spearman, C. E. (1927). The abilities of man. New York: Macmillan Company.

Special Issue on Writing Assessment. (1984). Educational Measurement: Issues and Practice, 3.

Standards for educational and psychological testing. (1985). Washington, DC: American Psychological Association, Inc.

Stiggins, R. J., Conklin, N. F. & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. Educational Measurement: Issues and Practice, 5, 5-17.

Stigler, S. M. (1986). The History of Statistics: The measurement of uncertainty. Cambridge, MA: Harvard University Press.

Thorndike, E. L. (1904). An introduction to the theory of mental and social measurements. New York: Teachers College, Columbia University.

Thorndike, E. L. (1910). Handwriting. Teachers College Record, 11, 83-175.

Thorndike, E. L. (1914). The measurement of ability in reading. Teachers College Record, 15, 207-277.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. In Whipple, G. M. (Ed.), The seventeenth yearbook of the national society for the study of education. Part II, The measurement of educational products. Bloomington, IL: Public School Publishing Company.

Thorndike, E. L. (1919). An introduction to the theory of mental and social measurements. Revised and enlarged edition. New York: Teachers College, Columbia University.

Thorndike, E. L. (1921). Measurement in education. Teachers College Record, 22, 371-379.

Thorndike, E.L., Bregman, E. O., Cobb, M. V. & Woodyard, E. (1926). The measurement of intelligence. New York: Bureau of Publications, Teachers College, Columbia University.

Traub, R. E. (Ed.). (1984). Special issue on the application of computers to educational measurement. Journal of Educational Measurement, 21.

Travers, R. M. W. (1983). How research has changed American schools: A history from 1840 to the present. Kalamazoo, MI: Mythos Press.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-482.

Wood, B.D. (1923). Measurement in higher education. Yonkers on Hudson, New York: World Book Company.

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

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

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. Journal of Educational Measurement, 23, 299-325.

Yen, W. M., Green, D. R. & Burket, G. R. (1987). Valid normative information from customized achievement tests. Educational Measurement: Issues and Practice, 6, 7-13.

Reference Notes

1. Support for this research was provided through a Spencer Fellowship from the National Academy of Education. An earlier version of this paper was presented at the annual meeting of the American Educational Research Association in New Orleans (April, 1988).

2. Although most of us are fairly familiar with E. L. Thorndike and his life, Ben Wood may not be as well known. Wood was involved, along with William S. Learned, in the Pennsylvania Study which was supported by the Carnegie Foundation from 1928-1932 (Resnick, 1982). One of the major outcomes of this study was to encourage high schools and colleges to keep cumulative records of their students. Wood also played a major role in the development of the Cooperative Test Service in 1930, as well as in the early development of the National Teacher Examination (Downey, 1965).

Table 1

Summary of the Principles of Educational Measurement

Principle	Problem	Proposed Solution
Objectivity	How can agreement on the meaning of a test be increased?	Development of a common measuring stick; reduce variation due to scoring method
Defined zero and unit	How can the location and unit of measurement be adequately defined?	Use Mean for location and SD for unit because of their relative stability
Definition of function to be measured	What is the test measuring?	Use clear operational definition of construct
Consistency	Is the test unidimensional?	Minimize obvious impurities
Within person variability	How can response errors due to intra-individual variability be minimized?	Increase number of items
Comparability	Can the test be validly applied with reasonable ease and accuracy to the objects to be measured?	Select an appropriate test to measure the construct

Table 1 (cont.)

Principle	Problem	Proposed Solution
Power and achievement are distinct	What is the difference between intelligence and achievement? Why is it important?	Difference must be kept in mind for interpretation and use of test results
Equal exposure and practice	How can differences in opportunity to learn be addressed when testing general intelligence?	Tests should be free of informational content; Take into account if control is not possible
Advantages of indirect measurement	What are the advantages of objective items as compared to essay items?	Use objective items to increase reliability
Test Construction*	What are the steps in test construction that can increase validity?	Increase number of items; use multiple item types; construct "good" items
Test use*	What are the steps in test use that can increase validity?	Appropriateness of the match between persons and items; reduce chance influences.
Measurement must not be confused with pedagogy	What is the distinction between measurement and pedagogy? Why is it important?	Treat testing and educating separately

* Note. Six principles treated separately by Wood have been grouped under test construction and test use.