2000

Current Trends in Psychological Testing

Kimberle L. S. Crawford
Eastern Illinois University

THESIS/FIELD EXPERIENCE PAPER
REPRODUCTION CERTIFICATE

TO: Graduate Degree Candidates (who have written formal theses)

SUBJECT: Permission to Reproduce Theses

The University Library is receiving a number of request from other institutions asking permission to reproduce dissertations for inclusion in their library holdings. Although no copyright laws are involved, we feel that professional courtesy demands that permission be obtained from the author before we allow these to be copied.

PLEASE SIGN ONE OF THE FOLLOWING STATEMENTS:

Booth Library of Eastern Illinois University has my permission to lend my thesis to a reputable college or university for the purpose of copying it for inclusion in that institution's library or research holdings.

_____ _____

Author's Signature Date

I respectfully request Booth Library of Eastern Illinois University **NOT** allow my thesis to be reproduced because:

_____

_____

_____

_____ 12-15-2000

Author's Signature Date

Title of thesis here

CURRENT TRENDS IN PSYCHOLOGICAL TESTING

Your name here

KIMBERLE L. S. CRAWFORD

1982 -

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

Master of Arts in Clinical Psychology

IN THE GRADUATE SCHOOL, EASTERN ILLINOIS UNIVERSITY
CHARLESTON, ILLINOIS

2000
YEAR

I HEREBY RECOMMEND THIS THESIS BE ACCEPTED AS FULFILLING
THIS PART OF THE GRADUATE DEGREE CITED ABOVE

12/15/2000
Date

Thesis Director

12/15/2000
Date

Department/School Head

Current Trends in Psychological Testing

Kimberle L. S. Crawford

Eastern Illinois University

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I want to fondly mention my thesis chair, Dr. Bill Kirk, and give heartfelt thanks and gratitude for his abundance of patience, calm instruction, and willingness to endure with me through the many hardships that arose during the completion of this writing. Without your conviction to guiding me in this project, it may have floundered.

Dr. Russell Gruber, thank you for your gentleness, encouragement and support when working with me. Your humor was always uplifting and somehow there at just the right time.

Most importantly I want to thank my father, mother, brother, sister and my pets for their unconditional love and faith in me. You have been most intimately aware of my struggles and it is your support, caring and best wishes for me that have given me the impetus to move ahead. I love you all very much.

To my husband Bill, thank you. You came through in the last of minutes and helped put the finishing touch on this. Smile!

To my close friends and loved ones, thanks for getting the coffee and chocolate out all those late nights and lending your emotional support.

ABSTRACT

Discussions about the adequacy of psychological measurement and assessment can quickly become controversial therefore; I expect some strong reaction to portions of this manuscript. Debates about the usefulness of criticism of psychological testing are longstanding: Even early psychologists such as Cattell and Jastrow disagreed on this issue. To be clear, I do not believe that use of contemporary tests should cease. I share the view that "psychological tests often provide the fairest and most accurate method of making important decisions" (K. R. Murphy & Davidshoffer, 1988, p. xii).

My first purpose, then, is to provide a historical survey of relevant measurement and assessment concepts. I do not delve into intimate details and complexities, but trace measurement and assessment controversies over time and across psychological domains. This approach produced a broad picture of how psychological measurement and assessment have evolved. The first half contains descriptions and interpretations of issues that have been important over the lifespan of psychological science.

My second goal is to expand discussion of the possible directions of measurement and assessment beyond those typically considered. The later half of this writing contains a summary of traditional approaches along with newer concepts and procedures.

It is important to expand the scope of topics typically presented in psychological measurement and assessment texts, and I offer this as a complement to those works. At the same time, I am trying to present this material as simply as

possible. Too much of measurement is incommunicable because of its complexity. My goal has been to approach the problems of measurement and assessment from the perspective of psychological theory. I hope to reconnect measurement with substantive theory to create, "better, richer, thicker descriptions from which to generate deeper conceptions and, most importantly, better questions" (Snow & Wiley, 1991, p. 6).

I have been surprised with the relative scarcity of sources describing the history of nonintellectual testing. Most measurement and assessment texts present a bit of history, and a few excellent book chapters and articles exist (Dahlstrom, 1985; Dawis, 1992). But I could find few sources that systematically examined the evolution of measurement and assessment to the extent that, for example, Boring (1957) did with experimental psychology. Perhaps the relative youth of psychological measurement—barely 100 years old—is a partial explanation. One consequence of this gap is that accounts of major concepts and procedures tend to be scattered throughout the literature. One of my goals has been to collect and reorganize seemingly unrelated material around long-standing measurement issues. At the same time, I expect that many readers will find portions of this repetitive or may find significant omissions.

HISTORY

BEGINNINGS OF CONFLICT

There is a tug of war going on between those that identify themselves primarily as scientists and those who identify themselves primarily as practitioners. The numbers and influence of practicing or applied psychologists whose specialty areas include clinical counseling, school and industrial or organizational psychology have grown dramatically during psychology's brief history. Fretz and Simon (1992) noted that the previous division between scientists and practitioners "seems to have become more of a chasm" (p. 31). Scientific psychologists' alarm at the practitioners' growing influence in the American Psychological Association (APA), the major United States psychological organization, led to the formation of an entirely new group, the American Psychological Society.

One of the most puzzling aspects of this split is the groups' inability to recognize that for both to thrive, they need one another (Danzinger, 1990). Throughout psychology's history, financial support for scientific psychology (typically housed in academic departments in colleges and universities) has often been provided on the premise that the work of such scientists would ultimately improve the lives of individuals. Similarly credibility for the interventions implemented by practicing psychologists has often been based on the belief that those interventions were determined through scientific methods (VanZandt, 1990). Dawis (1990) noted, for example, that early clinicians, armed with psychological tests, "had a technology for client assessment that had the decided appearance of professionalism...and the

scientific substance as well" (p. 11). Thus, despite their different goals, the science and practice of psychology complement one another. Most psychologists recognize the importance of meshing the two identities, as witnessed by the adoptions by some specialties of a scientist-practitioner training model in which graduate psychology students are trained in both research and practice skills (Barlow, Hayes, & Nelson, 1984; Gelso, 1979). That integration, however, has never been very successful: relatively few psychologist conduct research beyond that required in graduate school; relatively few express interest in psychological science jobs at the beginning or end of the graduate school career; and very few clinical practitioners base their work on research information (Barlow, 1981; Hermann, 1993).

Desire to be Scientific

In 1979 psychologists celebrated the 100th anniversary of the opening of Wundt's laboratory in Germany, an event often cited as the birth of scientific psychology. Organized scientific psychology is just over a century old. The fact that psychology is a young science must be included in any study of the history of psychological measurement.

Although psychology came to be more formally recognized as a scientific discipline in the late 1800s, events were taking place earlier in the century that shaped the early practices and procedures of scientific psychology. The most important events were related to developments in physiology, biology, and astronomy.

Model of Physiology. Success in research and measurement of physiological processes provided examples of ideas and techniques that psychologists could apply in their work. Wundt produced what some considered the first psychology book, *Physiological Psychology* (Heidbreder, 1933), describing how psychological experiments could be patterned after physiological ones: the scientist employs a controllable stimulus and records the objective responses of subjects. Likewise, Helmholtz provided practical demonstrations in his research on the eye and ear, demonstrating that the experimental methods and measurement techniques of physiology and the natural sciences could be applied to psychological phenomena (Heidbreder, 1933).

Another early psychological researcher, Fechner, viewed studies of just noticeable differences—for example, distinguishing between objects of slightly different weights—as revealing a mathematical relationship between the physical objects themselves and a person's perception of those objects. Fechner's goal was to apply the methods of physical measurement to psychological phenomena (Falmagne, 1992). Fechner believed that a person's ability to perceive a physical stimulus could be described by a mathematical function between the perceived sensation and measured stimulus value. With evidence of such a general relation, Fechner was strengthened to believe that psychological phenomena could be studied with the scientific method.

As the methods and measurements of the physiological laboratory became available to psychologists, sensation and perception appeared to be candidates for the basic elements from which all important psychological articles were constructed. Galton thought that sensory discrimination might be a sign of an individual's capacity for intelligent judgment. Cattell employed tasks such as grip strength, detecting the slightest differences in weight between two objects, and reaction time to sound in an attempt to develop predictors of intelligence.

Biology and individual differences. Darwin's publication in 1859 of *Origin of Species* provided another model for psychology. Two ideas seemed particularly relevant. First, individuals have to be considered in light of their abilities to adapt themselves to their environment. Second, humans passed on to their descendants a genetic history. Individual offspring displayed slight differences from their parents, differences that Darwin believed could be the source of materials for the processes of natural selection (Dawis, 1992).

Psychological experimenters have tended to assist individuals in finding the best environmental fit and to challenge agencies to adapt to the individual.

Darwin's cousin Galton became interested in the role of heredity on intelligence and developed testing methods that delved into the individuals personal and family history. Galton is thought of as the first to use items that required psychological ratings. His methods of measurement included tests of imagery where individuals were asked to recall a past experience in as much detail as possible. He

recognized differences in intelligence among individuals and used an approach that required tests to quantify the differences. Then a controversy in astronomy fueled interest in individual differences (Rosenthal, 1976). Nevil Maskelyne fired his assistant when his timing of transit stars across a certain point did not match his own recordings. It was soon found that individual differences in reaction time explained the different measurements not the assumed inadequacy of the assistant. Following scientists in other fields, Galton started a testing laboratory to assess individual differences (Danzinger, 1990). Soon other laboratories were opened and psychologists continued to demonstrate new tests and methods of collecting data (Boring, 1957). It was recognized that different individuals could behave differently on the same tasks or in the same situation.

Thus psychologists adopted some of the forms of the natural sciences. They emulated the physiologist's laboratory and experimental methods; they employed physical and physiological tests with which they could presumably measure psychological attributes. They adopted some of the current philosophical assumptions about the role of heredity, particularly heredity's role in determining intelligence. They emphasized the quantification of individual differences in psychological measurement. With these imitations, the young field of psychology could act scientifically and appear scientific to psychologists and the public.

The Desire to be Relevant

Although psychologists wished to be scientific, they also wanted to be socially relevant. As a profession, psychology has always been acutely aware of its social responsibilities.

Schools and wars. In response to the French government's need for a procedure to classify students as appropriate for formal schooling, Alfred Binet developed what came to be called an intelligence test. Binet may be credited with setting in motion a set of events that have had profound influence on measurement methods. Binet came to believe that intelligence was constructed not by simple elementary processes but by the whole of an individuals mental processes therefore, while being an experimental psychologist, he took the risk of abandoning the psychophysical tasks of the experimental laboratory. The Binet-Simon test excluded tasks with direct schoolroom content (Dahlstrom, 1985) but did include practical, academically similar tasks, like naming body parts and recalling pictures of familiar objects after viewing them for only 30 seconds. Since Binet was to determine student's likelihood of success in school, it is no surprise that the tests he used often resembled what they were supposed to predict (Frederiksen, 1986). Clearly, Binet's tests became the model for all psychological tests (Dawis, 1992).

The need to classify students wasn't limited to educational selection. The vocations movement began with Parsons 1909 publication, Choosing a Vocation (Shertzer & Stone, 1980). He believed that students needed systematic help in

choosing a career path. His method used a matching model in which student's abilities and interests were compared to the requirements of select vocations (Super, 1957). These traits could be measured through intelligence tests and questionnaires like Strong's Vocational Interest Blank (Strong, 1943).

During World War I and World War II, the military needed procedures for classifying the abilities of a large number of recruits to fit military occupations and to provide individual attention to soldiers in personnel selection, classification, and training (Dawis, 1992). Psychologists responded by adapting the tasks and procedures of Binet and others so they could be administered to large groups (K. R. Murphy & Davidshoffer, 1988). The Army Alpha and Beta intelligence tests, which resulted from these adaptations, were designed to screen out duller recruits and identify brighter ones for more responsible positions. Military needs did not center solely on selection issues. With the Korean War and World War II, there was a need for the development of a procedure to analyze the learning objectives for the training of soldiers in such tasks as assembling and disassembling and M-1 rifle (Bloom, Hastings & Madaus, 1971).

The efforts of applied psychologists in the first half of the twentieth century were largely devoted to solving pressing administrative and selection problems (Danzinger, 1990). Specifically, school and military administrators needed procedures for classifying large groups of individuals, the primary measurement needs of those administrators were, efficiency, obtaining as much information in as little time as

possible, and control, obtaining sufficient predictability so as to assign individuals to appropriate school programs, jobs and so on. Because of these needs, tests were designed to be as short as possible, measures of psychological traits present in all individuals, administered to large groups and evaluated primarily by their ability to predict important criteria. Psychologists later applied the same criteria to psychological tests that were not administrative in nature.

CONSEQUENCES OF THE SUCCESS OF PSYCHOLOGICAL MEASUREMENT

Breaking away from their roots in physiological and experimental methods, early measurement psychologists developed tests that met selection purposes in the first 50 years of the century. The methods and assumptions used during this time helped shape the future of psychological measurement. Psychologists today are trying to deal with the problems brought about by this original blueprint.

Prediction over Explanation

Binet's test did allow some prediction of school performance but little consensus was reached about how or why the tests worked. The test itself did not describe how an individual arrived at correct or incorrect answers. The emphasis was clearly on making predictions for selection decisions rather than understanding the meanings and causes of such scores. The search for causality and meaning was relegated to the background.

Binet and others used the construct of intelligence to address this issue. Basically, people who were more intelligent were more likely to answer the question correctly. Although it was generally assumed that the test and school criteria were indicators of intelligence, it was soon made apparent that no consensus could be reached about what constituted intelligence. To justify their efforts, mental testers adopted a philosophy that one could value a procedure without understanding it (Cronbach, 1992). It was thought that intelligence could be measured even if you didn't know what intelligence was. Intelligence was defined as what was measured by

an intelligence test. If the measurement procedure predicted something of interest, psychologists assumed that they would come to understand how it worked later. Today, psychologists can still avoid examining the meaning of what tests measure. In support of this position, Meehl (1954) argued that "item endorsements are interesting bits of verbal behavior whose nontest meanings remain to be determined" (p. 625). Epstein, (1979) observed that "Psychologists seem to take great pleasure in developing new scales..., but little interest in determining what scores on them mean" (p. 381). Early psychologists had little interest in knowing what specific tests actually measured, taking test labels at face value (Cronbach, 1992). Gould (1981) suggested that psychologists committed an error by trying to make something abstract concrete: They believed that because the term intelligence was implied to label what tests supposedly measured, intelligence came to have an independent existence of its own. Historically, psychologists seemed aware that other factors might be influencing scores on a test. Questions about and methods of investigating a test's validity—defined as whether a test measures what it is supposed to measure—did not become important until the 1950's.

The issue of test meaning, however, has critical implications. R. P. Martin (1988) noted that quite different social implications occur if one presents intelligence tests as partially flawed predictors of scholastic or occupational achievement as opposed to tests of one's stable capacity to learn. He suggested that if psychologists understand intelligence tests in terms of a stable capacity to learn, they will be inclined

to use those scores in a policy of selection. On the other hand, if intelligence tests measure culture-specific cognitive skills, abilities, or aptitudes, (Gronlund, 1985, p. 296) that result from individuals' unique learning history, psychologists will develop intervention programs to improve those skills. Historically, psychologists have assumed that intellectual ability is indicated by scores on intelligence tests, regardless of test-takers' cultural origins (Helms, 1992). He found that little research has been conducted to determine, for example, whether the ideas and concepts contained in intelligence test items are presented in a language equally understood by all cultural groups. Given that few measurement theorists have paid attention to the processes involved in test response, it is not surprising that researchers have failed to investigate the test-taking processes and strategies of different cultural groups.

The problem may partially result from the desire to be scientific to formulate general laws that hold across time and persons. Danzinger (1990) noted, "the early investigators in educational psychology were never content with a simple statement of practical expedience but always claimed that their statistical findings reflected on the nature of fundamental human functions or capacities, like memory or intelligence" (p. 150). Matarazzo (1992), for example, suggested the Binet "stumbled on classes of verbal and performance items that were included for the most part with what Spearman already had begun to identity as a "general intelligence factor". To qualify psychology as a science, psychologists desired laws that generalized as much as possible, laws about persons that went beyond specific individuals, time periods, or

situations. Intelligence tests were assumed to measure a phenomenon assumed to be inherited, stable across time, and the presumed cause of one's success in life or lack thereof.

Intelligence thus became the first important psychological trait, that is, a phenomenon assumed to be relatively stable, enduring, and resistant to environmental influences. What is crucial about a trait is its consistency wherever one looked. One should expect to find an individual's behavior consistent over such conditions as time and situations. The entire human population was assumed to be consistent in that psychologists assumed that traits were shared by all individuals. Most measurement psychologists have agreed that the first principle of measurement is reliability, defined as the consistency of measurement. The second major principle, valitidy, is commonly assumed to hinge upon the first: an unreliable test cannot be valid (K. R. Murphy & Davidshoffer, 1988). First and foremost scientists have searched for consistency. Once consistency was found, the search for universal laws could become relatively straightforward.

Despite progress, contemporary measurement authors continue to cite lack of theory as a major obstacle to developing better measurement and assessment instruments. There are hundreds of psychological tests but no analysis in terms of basic explanatory principles, with no methodology for producing that connection. Personality theories arise within naturalistic study—as in psychotherapy--or in the construction of psychometric instruments. Not only is there a need for more reliable

and valid personality tests, but improvements in the theoretical bases of these instruments and in the criteria against which they are validated are necessary...It is generally recognized, however, that none of the available methods of assessing personality is completely satisfactory. The solution to the problem clearly lies in better research and development (L. R. Aiken, 1989, p. 418). However, some sort of theory is implicit in every scaling approach and every test (Coombs, 1964). Supporters of the selection approach appear to assume that useful data are available only at the collective level but not with individual items, that traits are the major focus of the psychological measurement, and the criteria used to evaluate these had to relate to consistency and the test scores were directly related to what they were designed to predict. Clearly there are opportunities that allow us to look at measurement facts at the level of individual task or item, and in the contest of indirect relations (Gleick, 1987) between tests and criteria.

Loss of Experimental Methods

Binet demonstrated that school performance could be predicted without following traditional experimental techniques, thus, in the period that followed, there was an increase in testing that was not connected to traditional experimental questioning. With developments in sampling theory, results could be generalized from sample to whole populations (Heppner, Kivlighan, & Wampold, 1992). A shift occurred for experimenters which led their attention away from measurement concerns towards investigations on the effect of educational and therapeutic treatment. Thus

began a split between experimental and psychometric traditions in psychology that continues through the present (Cronbach, 1957, 1975b).

Experimentalists seem to treat individual differences as error, while psychometricians tend to ignore situational factors. (Cronbach, 1957; Danzinger, 1990) Experimentalists tend to look for the best intervention to apply to individuals in general while correlationalists look for the individual that would benefit the most from a treatment (Cronbach, 1957). Historically, experimental psychology has spent little time focusing on construct validity and its dependent variable while construct validity has become the focus of measurement validation efforts. These differences resulted in partial descriptions of psychological phenomena that have impeded the theory and measurement process in experimental and measurement psychology. There were those that preferred the experimental laboratory and those that were aligned with applied psychologists working in industry and education. In short, while experimental and psychometric work gradually separated, psychological measurement became closely associated with applied psychology. Heidbreder (1933) maintained that without experimental methods, measurement psychologists lost the capacity to study the factors that gave rise to individuals' performance on psychological tests. Clearly, it is thought that the loss of experimental methods hindered measurement progress.

Gain of Traits and Loss of Situations

Psychologists interpreted Binet's results as evidence of an intelligence factor.

Intelligence was considered to be a psychological trait thus, intelligence testing which came to be the model and standard for all psychological testing, emphasized the importance of enduring psychological attributes or traits over environmental influences.

Research has provided support for the importance of heredity in intelligence and temperament; however, many psychologists believe that situational and environmental influences play a major role and must be taken into account. This controversy has been mentioned more recently in terms of consistency of behavior across situations.

If psychological experiences are traits, then individuals who are truthful would exhibit truthfulness in all situations they encounter. Alternatively, some psychologists believe that psychological phenomena or traits would change depending on the environment or situation influencing the individual to be truthful in some situations but not in others.

Loss of Precision

Although Binet thought that his tasks were more valid that the physiological measures, he also believed that he had lost the precision of the laboratory tasks (Freeman, 1955). Precision has been described as the degree of exactness with which a quantity is measured. (J. T. Murphy, Hollon, Zitzewitz, & Smott, 1986).

Binet did sacrifice some precision. Psychophysiological tasks such as RT could reliable detect small differences. Binet indicated that his test would yield a

classification or hierarchy of intelligence that would be sufficient for practice. The practicing psychologist, then as now, needed a procedure that could produce gross, but still reliable differences between subjects who were from high or low on various abilities. With his tasks, psychologists possessed a reasonable, data-based procedure for selection purposes.

Many psychological tests can successfully place individuals into broad but distinct categories. We can be fairly certain that individuals that score low or high on a psychological characteristic really differ by a substantial amount for the measured characteristic and the they are likely to behave differently on some criterion that the test had been demonstrated to predict. We are more uncertain that the test could distinguish between the people who score in the middle range and the high or low groups. Unfortunately, "one feature which all psychological tests share in common is their limited precision" (K. R. Murphy & Davidshoffer, 1988, p. 2).

The important question about imprecision is, what causes it? It may be that imprecision is caused by the presence of multiple invalidities. Every psychological test is influenced by more than one factor; a score on psychological tests reflect several causes. If item responses are affected by more that one factor, then those responses are likely to change when the test is administered more than once. Some aspect of the test-taker or testing conditions is likely to be different from one occasion to the nest, thereby changing test responses over time.

CONSEQUENCES OF THE PROBLEMS OF PSYCHOLOGICAL MEASUREMENT

Even though there has been many successes in the field of psychological measurement and assessment, serious problems remain.

Scientific Crisis

Kuhn (1970) proposed that scientists in any discipline possess expectations about the results of their research. When these expectations are violated the attention of the scientific community focuses on the source of the violations. If the violations cannot be explained in a satisfactory manner, a crisis is precipitated. The failure to find expected consistencies in different measurement areas has caused recurring crises. Psychologists who work in mental health settings frequently deal with clients that are in crisis. These types of intense crises can not be maintained for the long term therefore, they must be resolved before the individual becomes physically and emotionally exhausted. Crises then refers to these meanings, measurement efforts have often violated the expectations of the scientific community and precipitated crises, which are later discarded or prematurely "solved" when the psychological community becomes exhausted and ceases its pursuit. The crises then reappears when the community recovers sufficient interest to renew the debate or when a new event indicates that the previous solution was insufficient. Meehl (1991) proposed a similar cycle for theories in the soft areas of psychology: (a) a new theory produces initial enthusiasm that (b) leads to considerable empirical investigation, but with ambiguous

results, followed by (c) a proliferation of hypothesis designed to fix unexpected results, and eventually leading to (d) abandonment of the theory and research as investigators lose interest.

The historical pattern in psychological measurement has been for an event to occur that ignites controversy, only to fade away as psychologists fail to reach a consensus. For example, personality psychologists have spent the past 25 years attempting to rebut Mischel's (1968) claim regarding the instability of personality traits. Although some psychologists now claim victory regarding the existence and importance of traits, (Epstein, 1979; Goldberg, 1993), it is unlikely that contemporary behavioral assessors would concede the potency of trait influence over situational factors. For most psychologists this issue evolved into a problem of the interaction of traits and situations, but what began as a promising solution to this problem has also run into difficulties (Cronbach, 1975;R. B. McCall, 1991). Such controversies are common in psychology. Recently, for example, rifts have developed regarding the adequacy of psychological assessment and diagnosis as it is presented by expert witnesses in legal testimony (Faust & Ziskin, 1988; see also Bulkley, 1992; Matazarro, 1990, 1991; Ziskin & Faust, 1991).

The likely consequence of repeated crises in areas related to psychological measurement are: (1) psychologists might neglect measurement and assessment in important ways, and (2) psychologists will become very cautious in their use of measurement procedures, over-relying on traditional tests and methods.

The Neglect of Measurement

Training. The neglect of measurement occurs in that psychology educators know that most students, particularly those in the applied specialties, begin graduate school with more interest and competence in qualitative matters than quantitative. Surprisingly, this situation changes very little by the end of graduate school, according to a recent survey of graduate departments of psychology (L. Aiken et al., 1990; see also Davison, Damarin, & Drasgow, 1986). About 25% of the departments rated their graduate students as skilled with psychometric methods and concepts: students were most knowledgeable about methods of reliability and validity measurement, less so with exploratory factor analysis and time analysis, and nearly ignorant of item-response and generalizability theories. Only 3% of the departments offered test construction courses. Howell (1992) wondered "how many of today's new doctorates in psychology really understand the psychometric underpinnings of the instruments they use or what they are doing when they-or their 'statistical experts'-apply a statistical package to a set of data" (p.21). Cliff (1992) maintained that "among all the variants of theory in our discipline, measurement must rank as the oldest in tradition, the deepest in formal development, but, for the majority of psychologists, the lowest in comprehensibility" (p. 88). Lambert (1991) believes that this situation constitutes a "crisis in measurement literacy" (p. 24).

Many studies have reported a tendency by researchers to be sloppy in their development of research scales and the reporting of scale characteristics. R. Hogan

and Nicholson (1988) stated that the literature is replete with examples of researchers testing substantive hypothesis with homemade and unvalidated scales; when it is later discovered that the scales did not measure what they purported to measure, the entire line of research is called into question (p. 622). Meier and Davis (1990) examined trends in the reporting of psychometric properties of scales employed over the previous three decades in the Journal of Counseling psychology. Although researchers increased the amount of reliability and validity data they reported over the sampled periods, the majority of scales were still accompanied by no estimates. Meier and Davis (1990) also found that 1/3 of all scales were either investigator developed for the special study or were modified by the investigator. Reliability estimates that were reported provided some evidence that little progress has occurred in psychological measurement. Babor, Stephens, and Marlatt (1987) found similar infrequent reporting of scales' reliability and validity estimates in the Journal of Studies on Alcohol. Tinsley and Irelan (1989) stated, "It is clear that the emerging empirical base upon which the profession is being built is founded to a substantial degree on instruments that have enjoyed little or no rigorous scientific scrutiny" (p. 446).

Wisdom and Absolutism of Tradition

Psychologists appear to ignore problems with a measurement procedure if the device fulfills some purpose. Despite our lack of understanding of what those tests actually measure, psychologists have long employed intelligence tests because of the

test' predictive capacities. Although they admit that it does not posses the predictive validity of intelligence tests, MMPI proponents point to the 10, 000 published studies (Graham, 1990) as evidence that we know something about the meaning of its scales. In the context of repeated scientific crisis that go unresolved, considerable safety exists in the use of a traditional test, particularly if no viable or distinct alternative exists. Surveys of clinicians find that they continue to rely on traditional measures such as the MMPI and the clinical interview (Lubin, Larsen, & Matarazzo, 1984; K.R. Murphy & Davidshoffer, 1988; Piotrowski & Keller, 1984, 1989; Piotrowski & Lubin, 1990; Frauenhoffer, 1995). When they create new tests, researchers often imitate previous measures. For example, developers of new intelligence tests have even tended to borrow the same items from previous tests. Jackson (1992) reported that Wechsler used items from the Army Beta in the Wechsler-Bellevue intelligence test. Hathaway (1972) noted that psychologists continue to employ 1940's technology in the use of the MMPI; it took nearly 50 years before a revision of the MMPI was accomplished. Similarly, Gynther and Green (1982) suggested that traditional self-report methodology has advanced no further since the scales of the 1940's and 1950's. Buros (1970) suggested that "it is sobering to think that our most widely used instruments in personality assessment were published 20, 30, 40 or even more years ago" (p. xxv).

Moderate success, repeated crises, and a lack of alternatives may motivate psychologists to stick with traditional instruments. But the side effect is that

psychologists tend to minimize the problems of the original device and fail to explore and experiment with new approaches that represent potential improvements. As S. Kaplan (1964) noted, "A conspicuously successful technique in some area of behavioral science is not only identified with 'scientific method,' but comes to be so mechanically applied that it undermines that very spirit of scientific inquiry" (p. 29). If problems do become so severe as to force abandonment of a particular measurement procedure, psychologists have tended to recycle the procedure again at a later date. Interviews, for example, have been employed throughout psychology's history, but they fell into disfavor when research suggested that interview data could be influenced by numerous factors, such as interviewer race and gender (K. R. Murphy & Davidshoffer, 1988) and that interviews were more reactive than questionnaires (Stone, 1978). Goldstein and Hersen (1990) recently observed, however, that "Following a period when the use of the interview was eschewed by many psychologists, it has made a return. It would appear that the field is in a historical spiral, with various methods leaving and returning at different levels" (p.4).

MEASUREMENT AND THEORETICAL ADVANCES

Measurement problems are noteworthy because of measurement's crucial role in fostering scientific progress. As Tryon (1991) stated, "the history of science is largely coexistent with the history or measurement" (p. 1). Cone (1988) agreed: "It is certainly beyond argument that the building of all science rests of a foundation of accurate measurement" (p.42). Even a quick review of the history of science indicates

that new measurement techniques drive scientific development (Cone & Foster, 1992; Forbes & Dijksterhuis, 1963). Meehl (1991) provided several examples: in chemistry, spectroscopy make possible knowledge about the composition of the stars; in biology, refinement of methods of measuring adenine and thymine along with advancements in x-ray technology make possible Watson and Crick's discovery of the structure of DNA. Scientists note that scientific progress is not solely dependent upon measurement advances, but new measurement techniques have allowed refined tests between real theories and hypothesis (B. Ellis, 1967). Continuous innovation in measurement theory and techniques seems to be important to progress in any science (Judson, 1980; Kuhn, 1970). Tyron (1991) indicated that this progress is enabled by measurement's capacity to provide new data for ideas and correct for limitations of the senses. Examples of the link between measurement and scientific progress can be seen in astronomy. Galileo did not invent the telescope, but he employed it to observe Jupiter and its revolving moons, thus setting the stage for acceptance of a heliocentric view of the solar system. Galileo also encountered a validity problem similar to measurement psychologists. Forbes & Dijksterhuis (1963) wrote, "No one yet understood the operation of the telescope and it is doubtful whether Galileo did so himself. In any case he made no attempt to prove that what he saw in his telescope really existed" (p. 210). He could make basic observations and predictions with his telescope, but he could provide little evidence to verify the truthfulness of what he saw. Progress in astronomical observations has been extremely rapid in the past

twenty years. Using the example of observational advances of Saturn, in the 17th century, observers, such as Galileo debated whether the objects surrounding Saturn were rings. Later scientists on the Mount Wilson Observatory were able to photograph clearly the division of rings. Then even later, photographs showed in clear detail the ring system surrounding Saturn. With psychological measurement, we are able to observe the surface structure, but very little detail.

Traditional views of the research process typically describe a cycle whereby theory produces inquiry and the resulting data refine and change theory. Platt (1977) described the particular process as follows: a) formulate two or more competing hypothesis, (b) devise a study that unambiguously contrasts those hypothesis, (c) run a clean study, one that is free of methodological explanations for the results, (d) use the results to reject unsupported hypotheses, (e) recycle the procedure. Following the procedure is expected to produce a conclusion with few or no alternative explanations. Research data, however, depend on the quality or lack of quality of the measurement procedure used to obtain it. If there are considerable measurement problems, it will be impossible to rerun a clean study. What it does do is present an alternative explanation. It is necessary to have a lack of alternate explanations in order for psychologists to proceed with confidence. The presence or absence of any type of alternative explanation, not whether a study is experimental or correlational is what ultimately is the basis for evaluating the helpfulness of a theory-research program.

Measurement problems pose a complication in the process of challenging and disconfirming psychological theory. Meehl (1967), Kazdin (1980) and Dar (1987) observed that "a common tactic among researchers is to lament the lack of sensitivity of a dependent measure as a possible explanation for the lack of findings" (p. 221). The investigator suspends judgment on experimental hypothesis by rejecting empirical results on the basis of suspected measurement problems. If measurement problems are substantial, this may be a correct decision. Sometimes the decision may result in a missed opportunity to change or negate a theory. Given the numerous and complex problems in psychological measurement, alternative explanations are plausible for many studies employing psychological measures. In experimental research in psychology, measurement scales typically form the dependent variables. In correlational research, the independent and dependent variable are often measurement scales. Substantial measurement problems, then, are likely to seriously impede the accumulation of knowledge in psychological research. This accumulation of knowledge ultimately depends on the ability of measurement devices to provide meaningful information.

Knowledge Accumulation. Many psychologists believer that the field suffers from a lack of a substantial knowledge base. Danzinger (1990) maintained that soft psychology has always had to contend with the problem that its knowledge base did not substantially diverge from common sense and that soft psychology has yet to develop expert knowledge in many areas. Gardner, (in Holder, 1988) suggested that "

the areas of personality, consciousness, self, and will...are scarcely more developed now that in the days of James and Freud' (p. 1036).

Compared to all psychological measurement devices, intelligence tests typically possess the best psychometric properties (K. R. Murphy & Davidshoffer, 1988). Yet the ability of those tests to make predictions has not substantially changed during the past 70 years. Jackson (1992) noted that the validity of psychological tests has not substantially improved over the Army Alpha and Beta tests. Observing that the Scholastic Aptitude Test (SAT) predicts only about 10 to 15 % of first-year college grade point average (GPA) variance, Wainer (1993) suggested that "predictive validity is too low and must be increased" (p. 2).

Kuhn (1970) predicted that once a scientific crisis is recognized, traditional rules will be loosened and more radical methods will be proposed to solve the problem. Such a situation certainly seems to exist in areas of psychology that suffer from a "relatively unimpressive degree of cumulative knowledge" (Feldman, 1971, p. 86).

Theoretical Accuracy. In addition to depth of knowledge, theory can be categorized as to its accuracy in describing natural phenomena. Accuracy refers to the degree to which a measurement approaches the actual value of the psychological phenomena measured (Kendall & Buckland, 1957, cited in May, 1979). High accuracy is ascribed when a theoretical prediction is shown to be true. High probability is a fundamental feature of strong theories (Green, 1992); such theories

make highly accurate predictions. Green (1992) also observed that the accuracy of strong theories are at least a partial result of theoretical precision. Precise theories in turn depend on data produced by precise measurement procedures.

Disunity. One of the important theoretical activities in any science is the making of conceptual connections between seemingly disparate phenomena. When we begin to explore any phenomena in depth, a more complete picture begins to emerge, particularly in regard to how events and characteristics that appeared unrelated do, in fact, have important connections. Zigler and Glick (1988), for example, proposed that schizophrenia is not a homogeneous disorder, but consists of distinct subtypes. They suggested that paranoid schizophrenia may be more accurately described as a camouflaged depression. If this connection between a type of schizophrenia and depression is valid, it has important implications for the treatment of paranoid schizophrenics.

Despite a wealth of psychological data and constructs, there exists "a rather serious shortage of important connections" (Torgerson, 1958, p. 5). Staats (1981) described psychology as separatist, meaning that the science is "split into unorganized bits and pieces, along many dimensions...Our field is constructed of small islands of knowledge organized in ways that make no connection with the many other existing islands of knowledge" (p. 239). Staats maintains that schisms, such as the role of personality versus environment in determining behavior, abound in psychology. Even a cursory review of various psychological literatures reveals phenomena that

investigators consider separate but that contain highly overlapping theory and methods.

Although Staats (1981, 1983) believes that the separatism of psychology is largely a result of the failure to seek unity, it may be that the field as a whole may not yet possess the capacity to make important connections. We may not yet have the right sort of data to make integrative efforts. What we need is more precise data to create theory with finer distinctions.

Decisions made in the beginning of a science influence subsequent trends and events (Danzinger, 1990). The concept of a trait was the primary unit adopted by early measurement psychologists. Although process and change were recognized in psychological phenomena, stable, enduring traits best fit early psychologists' assumptions about human nature and their methods of measurement. Consistency was understandable; inconsistency was error.

PROMISING APPROACHES

This chapter contains descriptions of important existing and innovative approaches in psychological measurement and assessment. I discuss theory and methods in terms of four major categories: traditional approaches (e.g., MMPI), cognitive approaches, behavioral assessment, and computer-based methods.

TRADITIONAL APPROACHES

Traditional measurement devices such as the WAIS-R, Rorschach, and MMPI-2 continue to be widely used by psychologists to assist in selection decisions in education, mental health, medicine, business and industry, and the legal system. As many authors have noted (Buros, 1970; Gynther & Green, 1982; Hathaway, 1972; R. P. Martin, 1988; K. R. Murphy & Davidshoffer, 1988), few innovations introduced since the 1940's have been powerful enough to alter the use of traditional tests and procedures employed by psychological practitioners and researchers.

Traditional tests share certain methods, concepts, and statistical assumptions. In classical test theory, an observed test score is composed of true score and error. The true score usually represents a trait, a relatively enduring personal characteristic that influences behavior across settings. The goal of classical test theory is to maximize the true score component and minimize error. During test construction and evaluation, classical measurement approaches attempt to identify and minimize error through statistical methods. Typically, self-report items completed by many individuals are aggregated to produce an estimate intended to discriminate traits and individuals from other traits and individuals.

MEASUREMENT OF PERSONALITY AND TEMPERAMENT

The Big Five

Given the desire for a classification of important traits, personality psychologists have reached a consensus about what are termed the Big Five factors (Cattell, 1946; Digman, 1990; Goldberg, 1990; McCrae & Costa, 1989; Norman, 1963; Wiggins & Pincus, 1989). These independent factors –neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness—have been proposed as singular structures to guide the measurement of personality and interpersonal behavior. Factor analyses of trait descriptors, produced by different methods (e.g., self-report and ratings by others) and with different samples (including cross-cultural), have resulted in the identification of five factors (Botwin & Buss, 1989; John, Anglietner & Ostendorf, 1988; McCrae & Costa, 1985). Two major components of the circumplex model of interpersonal behavior, dominance and nurturance have also been connected to the Big Five factors of extraversion and agreeableness (Trapnell & Wiggins, 1990).

Empirical results, however, provide some ambiguity about the Big Five model. Botwin and Buss (1989), for example, instructed 59 couples to report self and other data about previously performed behaviors corresponding to the five-factor model. Factor analyses of self and partner data yielded similar results. Botwin and Buss concluded that the resulting factors, labeled responsible-stable, insecure, antagonistic-boorish, sociable and culture, departed substantially from the five-factor model. When

ratings data were adjusted for the frequency level of the behaviors; however, the resulting factors closely matched the five-factor model.

Many traditional approaches grew out of a history of testing aimed at identifying stable traits for selection purposes. Except when political and legal forces intervene, traditional approaches, given their demonstrated effectiveness and fairness, are likely to be increasingly employed in the future for such decisions. For example, if pressure to reduce health care costs continues to shrink mental health benefits, it would not be surprising to see traditional tests such as the MMPI used to screen individuals to determine the degree of psychological disturbance and thus decide the amount or type of treatment they subsequently receive. I expect traditional tests to be employed less frequently in the future for guiding psychological interventions.

COGNITIVE APPROACHES

Scaling refers to "the assignment of meaning to a set of numbers derived from an assessment instrument" (Reckase, 1990, p. 41). The important issue here is who assigns the meaning to the numbers. Traditionally, the chief decision-maker is the test developer who transforms the raw data provided by test-takers using statistical methods. For example, the developer may (a) sum all item responses to produce a total score, (b) examine correlations between individual items and summed scores to determine which items should be deleted from a scale because of low correlations, or (c) factor analyze the data to determine which items converge to or from separate

constructs. In each case, the developer assures that the transformed data have more meaning than any particular response generated by the test taker.

In addition to the test developer, the test-taker or observer assigns meaning to numbers. Test items and tasks are cognitively and affectivity processed for meaning and subsequent response in a different manner by different individuals (Cronbach, 1946). As Loevinger (1957) stated, "the naïve assumption that the trait measured by a personality test item is always related in a simple, direct way to the apparent content of the item has long since been disproven" (p.651). This approach represents movement away from viewing questions and answers in a stimulus-response mode toward viewing them in terms of meaning and context (Mishler, 1986). The importance of understanding the processes employed by individuals when responding to tests has long been understood; for example, Cronbach (1946) cited work which indicated that test results are not simply a product of individuals' ability (or personality), but also of the methods subjects employ when completing test items. This knowledge, however, has had relatively little impact on psychological testing (Burisch, 1984).

Many contemporary psychologists have returned their attention to response processes (Guion, 1977). Experimental methodologies and theoretical constructs from cognitive psychology have begun to provide paradigms for investigating such processes in cognitive ability and other tests (Embretson, 1985). Embretson,

Schneider, and Roth (1986) suggested that aptitude researchers have moved their focus from traits to the cognitive components underlying test performance.

Cognitive approaches' great promise would seem to lie in their potential for illuminating important cognitive components that partially form the basis for construct validity. That is, cognitive approaches offer tools for investigating such problems as how test-takers construe the meaning of items and how individuals' cognitive abilities interact with test item characteristics. Historically, motivational and affective influences have also been recognized as important influences in testing processes, and it is here that cognitive psychology may have less to offer in the way of theory and methods.

BEHAVIORAL APPROACHES

Behaviorists and behavioral approaches emphasize the dominance of environmental reinforcement in shaping individual's behavior, be it motor or verbal. In contrast with traditional psychological measurement, behavioral assessments measure individuals' past learning histories and current environmental influences (R. O. Nelson & Hayes, 1986). Behavior is observed in natural or contrived settings and assessors attend to stimuli, behavioral responses, and the consequences of those responses. The processes, assumptions, and procedures of behavioral assessment differ from traditional measurement. Hartman (1984) emphasized that behavioral assessment is direct, repeated, and idiographic. Assessment is direct in that the psychologist measures observable behavior. Any observed behavior is considered to

be an example of potential behavior, as opposed to a sign of an underlying, unobservable trait. Behavior is measured repeatedly for the purpose of demonstrating relative stability before intervention and change after intervention, thus demonstrating that the intervention is the cause of the behavioral change. Assessment may consist of continuous recording of behavior (when only a few behaviors occur) of some type of time sampling. With the exception of areas driven by accountability concerns (e.g. psychiatric inpatients), nonbehavioral psychologists typically do little or no formal measurement during the intervention process.

Behavioral approaches generally assess variables that are unique to the individual in question, such as a behavior, affect, and cognition sequence. Behavioral assessment is typically performed in a clinical setting, in conjunction with behavioral interventions. In contrast with traditional psychological measurement, where anyone can be a self or other-observer if enough measurements are gathered to decrease measurement error, behavioral assessment involves trained observers.

Behavioral approaches offer well-developed methods for observing psychological phenomena. To the extent that behavioral assessment retains its radical roots, however, it is likely to resist exploring the usefulness of well-developed traditional concepts such as constructs, reliability, and validity. Such concepts seem indispensable for theory development and measurement evaluation.

COMPUTER-BASED APPROACHES

Scientific progress is inseparably linked to the state of measurement theory and procedures. Measurement, in turn, is limited by the technologies available to gather date. Throughout most of psychology's history, the predominant measurement technologies have been printed materials and pencils. Danzinger (1990) maintained that other sciences came to rely on reliable witnesses as the key to credible knowledge; nonetheless, technology can also increase observer reliability (Rosenthal, 1976). The moons of Jupiter, for example, are invisible to all except individuals with exceptional eyesight in excellent atmospheric conditions. A telescope, however, allows any sighted person to easily view those moons at leisure.

The introduction of microcomputers has resulted in widespread interest in measurement uses of this technology. With the exception or IRT applications, most contemporary developers of computer-based testing procedures have focused on adapting traditional tests so that one or all test components—administration, response recording, scoring and data analysis, and interpretation—is done by computer (Hedlund, Vieweg & Cho, 1984). Many of these applications were published in the 1980s when it appeared that testing software would be an economic boon for developers and publishers. Automated procedures were created, for example for the MMPI (Anderson, 1987; Honaker, 1988), 16PF (Harrell & Lombardo, 1984), Rorschach (Exner, 1987), Strong Interest Inventory (Hansen, 1987; Vansickle, Kimmel & Kapes, 1989), Self-Directed Search (Reardon & Loughead, 1988),

interviews, (Erdman, Klein & Greist, 1985; Giannetti, 1987) intelligence and aptitude tests (Harrell, Honaker, Hetu, & Oberwager, 1987), psychophysiological research (Blumenthal & Cooper, 1990) and diagnostic procedures (Stein, 1987).

Given that the basic objective of much of this work has been the transfer of paper-and-pencil tests to computer, a logical research question to ask concerns the equivalence of procedures, particularly with computer administration of test material (Skinner & Pakula, 1986). That is, does the automation of test procedures affect the instrument's reliability and validity? Given the economic potential of testing software, these questions have tended to be investigated after software has been developed and marketed. Some studies have found no differences between traditional and computer-administered versions of tests (e.g., Calvert & Waterfall, 1982). However, some who take computer-administered tests show elevated negative affect scores, indicate more anxiety with computer-based procedures (Hedl, O'Neil, & Hansen, 1973), alter their rate of omitting items, and increase their "faking good" responses (Davis & Cowles, 1989). Given the equivocal findings, the equivalence issue currently must be addressed by test administrators on a test-by test, sample-by-sample basis.

Although straightforward automation of traditional tests often improves the reliability and efficiency of testing procedures and scoring, computerization has yet to advance basic measurement theory and technology. The existing and growing base of microcomputers, however, offers a platform from which to support a second phase of

new measurement procedures that more fully utilize computer capabilities. Experimental procedures and measurements that have previously been laboratory-based can now be economically transported to microcomputers for use in applied settings. As Embretson (1992) pointed out, tightly controlled experimental tasks may be implemented as test items. Computer-based measurement can blur the distinction between experiments and tests, thus facilitating the unification of correlational and experimental psychology suggested by Cronbach (1957). Automation may make derivations possible in other domains, several of which are described below.

Response Latency

A variable usually associated with cognitive investigations in laboratory setting, reaction time (RT) or response latency can easily be measure in computer-based tests and tasks (Ryman, Naitoh & Englund, 1984). Brebner and Welford (1980) observed that early psychologists hoped to use RT as a physical measure of mental processes. Contemporary psychologists have employed latency as indicators of cognitive ability (Lohman, 1989; Jensen, 1982), stress and fatigue (Nettlebeck, 1973), and psychopathology (Nettleback, 1980).

Utilizing latency as a key component, R.R. Holden, Kroner, Fekken and Popham (1993) described a model to predict faking on personality test item response. In this model, test-takers respond to items by comparing test item content with self-information contained in a schema (R. R. Holden, Fekken & Cotton, 1991). Because schemas expedite the search for information, R. R. Holden et al. (1992) proposed that

responses should be faster for schema-congruent test answers than incongruent responses. Previous research has found that individuals who possess high total scores on an anxiety scale respond more quickly when agreeing with anxiety-relevant items (R. R. Holden et al., 1991; S. M. Popham & Holden, 1990). Given the historical emphasis on distortion in self-report, R. R. Holden et al. (1992) extended this model to include dissimulation on personality test items. They reasoned that persons faking good would respond more quickly to socially desirable items (i.e. congruent schema) than undesirable items. Conversely, persons, faking bad should respond more quickly to undesirable items than desirable ones. Using microcomputer-presented item, R. R. Holden et al. (1992) found support for both hypotheses in a series of studies utilizing the MMPI and Basic Personality Inventory with college students and maximum-security prisoners. Other researchers (M. George & Skinner, 1990; Tetrick, 1989) have also described studies using subjects' response latency to individual questionnaire items to detect inaccurate responding.

Human Speech

As applications to record and transcribe human speech become available and economical in this decade, they have the potential to revolutionize interviewing and measurement procedures. A program that could recognize disruptions of normal speech patterns and relate that information to anxiety (Mahl, 1987) would certainly be of interest to research and applied psychologists. Most theories of counseling and psychotherapy view language as crucial to understanding and intervening with clients.

Researchers and practitioners have considerable interest in computer programs that could transcribe psychotherapy sessions and produce or assist in the production of qualitative and quantitative measurements of that communication.

Research on speed applications appears to be in its initial stages. Friedman and Sanders (1992) described a microcomputer system, coupled with a telephone, designed to monitor pauses in speech. They analyzed long speech pauses (defined as > or = to 1 second) in relation to mood disorders. Friedman and Sanders (1992) maintained that pause measurement could be useful in studying and identifying such problems as depression, mania, dementia, and coronary prone behavior. Canfield, Walker, and Brown (1991) described a microcomputer-based coding system to analyze sequential interactions that occur in psychotherapy. Using the Gloria films of Ellis, Perls, and Rogers (Shostrum, 1966), Canfield et al. (1991) explored whether this coding system could demonstrate differences among therapists with distinct therapeutic styles. Canfield et al. (1991) analyzed transcripts of the Gloria psychotherapy sessions for positive and negative emotion, cognition, and contracts (i.e., promises and commitments). As expected they found that the three therapists' use of these categories differed in frequency. They also found that Gloria differed in her frequency of these categories across the three therapists and that therapist employed different sequences of categories when responding to client statements. For example, one therapist would response to a client's positive emotion statement with a positive emotion, whereas another therapist would respond with a positive cognition.

Canfield et al. (1991) noted that previous studies of these films have found differences in use of predicates (Meara, Shannon, & Pepinsky, 1979), reflection and direction (Hill, 1978), and language structure (Zimmer & Cowles, 1972).

Simulations

Computers make increasingly realistic simulations of the type discussed above an economically viable possibility now and in the near future. Multimedia programs that utilize audio and visual material in addition to text may be used to create assessment simulations for use in business, industrial, educational and clinical settings. These simulations can also function as unobtrusive measures to supplement reactive self report scales (Johnson, Hickson, Fetter, & Reichenbach, 1987). Computer-assisted instruction programs (CAI) can employ simulations to perform the dual functions of teaching and assessment (Fulton, Larson, & Worthy, 1983; Meier & Wick, 1991). Meier and Wick (1991) described a simulation designed to demonstrate blood alcohol levels for subject-selected drinking experiences. Unobtrusive recorder reports of subjects' alcohol consumption in the simulation was (a) significantly correlated with self reports of recent drinking behavior, drinking intentions, and attitudes toward alcohol, and (b) uncorrelated with a measure of social desirability. Similarly, Worthen, Borg, and White (1993) discusses the use of computers in continuous measurement in educational settings. If a particular curriculum was computer based, testing could be embedded in the instructional material and thus be relatively unobtrusive. Worthen et al. noted that such an approach fits very well with

mastery learning where progression to the next level of instruction depends upon demonstration of successful performance on the current material.

Being relatively resource poor, psychology usually must wait for new technology to become available in the mass market place before such devices can be applied to measurement and assessment problems. Such is the case with virtual reality, a set of computer-based devices that allow simulations of visual, auditory, and tactile phenomena. According to Potts (1992), the devices typically include (a) a helmet for projecting three-dimensional views and stereo sound; (b) a joystick for controlling the user's movement in the virtual world (c) a glove that allows the user to manipulate objects in the virtual world; (d) a Polhemus sensor, suspended above the user, that tracks the positions of the helmet, joystick, and glove, and relays that information to the computer; and (e) a computer to create sensory input for the user and track the user's actions. If the validity of simulations depends upon the closeness of their match to real situations (Motowidlo, et al., 1990), then virtual reality holds great potential for psychological measurement. Like much of the technology described in this section, however, the cost of a virtual reality system is high (down from $200,000 a few years ago to $20,000 currently, according to Potts, 1992) and availability is low.

I suspect computer-based approaches will be increasingly employed in the future for no other reason than they offer increased efficiency in test development, administration, scoring, and interpretation. Automation in psychological testing has

occurred, however, with relatively little attention to such theoretical considerations as human factors issues or fully employing computer capabilities in the testing process.

This chapter has described current work in traditional measurement, cognitive approaches, behavioral assessment, and computer-based approaches. Much of the work in traditional measurement appears to center on confirming the Big Five factor of personality and extending this model to other areas such as clinical assessment. Behavioral assessment continues to thrive even with a greater focus on it relations to traditional psychometric concepts. Cognitive theory and procedures hold considerable promise for the investigations and explanation of such measurement processes as item response; cognitive models, however, tend to neglect the motivational and affective influences of measurement processes. Computer-based approaches would seem important if for no other reason than technological innovation tend to reduce the amount of interpretation in measurement and assessment. Content analysis and qualitative research, for example, are facilitated by tape recording of conversations, which allows the listener to replay phrases and sentences for coding instead of performing the task as the activity occurs.

DISCUSSION

Accepting any successful paradigm carries with it benefits and costs. The promoters of a successful paradigm will obviously emphasize its benefits, but eventually other groups will attempt to demonstrate and publicize the accompanying costs. To the extent that a paradigm cannot explain the phenomena of interest, calls for revisions and modifications, and revolutions are likely to be increasingly heeded.

A paradox exists between the sharp criticisms of psychological measurement by academic scientists and others and the widespread use and acceptance of current tests by practitioners and researchers. The paradox may be understood by noting the psychological measurement devices appear adequate for some purposes, but not for others. For example, selection testing is one of psychology's most important societal functions. In education, business, and the military, psychological measurement, particularly cognitive ability tests represent the best and fairest approach to selection decisions. That success led psychologists to apply selection assumptions and procedures to tests in other areas, such as explanatory theory building and intervention, where the results have been less successful. Thus, ample reason exists both to respect contemporary tests and to work toward their improvement.

There are some that believe or imply that important psychological constructs are largely immeasurable (L. Goldman, 1990; Mayer, 1966) The evidence does not support Faust and Ziskin's (1988) assertion that human behavior" resists objective direct, or reliable observation and measurement" (p. 33). No other set of procedures

exceeds the utility of psychological measurement and assessment in their proper applications. To improve measurement's effectiveness, what is required is a renewed emphasis on theoretical and technological development, application of existing substantive theories to measurement procedures, experimentation with new procedures, and programmatic research. Increased funding, additional attention to measurement and assessment in graduate course work, and greater attention to these issues in scholarly publications should support these goals.

The issue of sufficient resources to study measurement and to do applied assessment is akin to putting the "cart before the horse" problem. Without additional resources necessary to improve measurement and assessment abilities, little progress in these areas can be expected. Yet those who judge our requests for additional resources in science and practice will expect better measurement and assessment. Economics is likely to be a critical factor in applied measurement innovations: for example, as long as costs are the primary factor in deciding the extent to which assessment should be conducted in practice settings, the most inexpensive methods, self-reports and interview, will continue to reign. It is ironic that the managed companies now dominating psychological practice, who publicly profess to value solid evaluations of health care, show so little interest in paying for psychological assessment, research on assessment innovation, or training practitioners in effective assessment.

Dawis, (1987) maintained that "researchers should not be reluctant to experiment with different scale construction approaches—and should report their results, so that the rest of us can find out what method is best"(p. 488).
Sidman (1960) observed that much important work in experimental psychology is devoted to "... improvements in measuring instruments, advance methods of recording date, sophisticated data analysis, the design of specialized apparatus to do a particular job or generalized apparatus to perform many functions, and the extension of old techniques to new areas" (p. 16).

When measurement and assessment advance, my thought is that testing will move away from its strict reliance on self-report. For example, several researchers have begun to investigate the potential benefits of combining self reports with the response latencies of those reports (R. R. Holden et al., 1992). Test construction will focus not just on maximizing individual differences, but on other criteria as well. I believe that scale development will become more specialized and closely linked with measurement theories.

Kuhn (1970) implied that one essentially gambles when choosing between paradigms. My bets are on the theories and procedures, such as the cognitive approaches that pay more attention to construct validity and construct explanation. Cognitive and cognitive-behavioral theories are now available that have the capacity of guiding investigations at the item level.

Although students still require instruction on such basics as reliability, validity, norms and item analysis, recent developments argue for a revision and possible expansion of the measurement curriculum. I would argue for greater attention to approaches that emphasize construct validity and construct explanation. To diminish the artificial science-practice split in the graduate curriculum, faculty should also increase the measurement emphasis in applied and practice courses. Instructors could construct assignments in which students apply basic and advanced measurement procedures to important problems in their subspecialties. Such practical application might increase student's interest in measurement and assessment and pique some student sufficiently to investigate the utility of new measurement procedures in their areas. Few students realize that the investigator developed scales frequently employed in basic research (Meier & Davis, 1990) lack sufficient evaluation in terms of reliability and validity, thus presenting alternative methodological explanations that weaken confidence in research results. In general, measurement instructors should teach students how to evaluate the psychometric properties of measurement devices employed in published studies and routinely evaluate and report the psychometric properties of students' measurement devices.

Throughout the history of the science and the profession, psychologists have been presented with a pressing need for useful measurement and assessment tools. The practice of measurement and assessment will continue, and without substantial improvement, so will the crisis. Faust and Ziskin's (1988) controversy about the

adequacy of psychological assessment and diagnosis as presented by expert witnesses in legal testimony is now moving away from center stage in favor of questions about clinicians and their clients abilities to distinguish between repressed memories and false ones (Loftus, 1993). I believe that significant progress in measurement and assessment is the single most important prerequisite for unification of psychological science and practice.

REFERENCES

Aiken, L. R. (1989) Assessment of personality. Boston: Allyn & Bacon.

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Rodiger, H.L., III, Scarr, S., Kazdin, A. E., & Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of Ph.D. programs in North America. American Psychologist, 45, 721-734.

Anderson, R. V. (1987). Computerization of a chemical dependency assessment. Minnesota Medicine, 70, 697-699.

Babor, T. F., Stephens, R. S., & Marlatt, G. A. (1987). Verbal report methods in clinical research on alcoholism: response bias and its minimization. Journal of Studies on Alcohol, 48, 410-424.

Barlow, D. H. (1981). On the relation of clinical research to clinical practice: Current issues, new directions. Journal of Consulting and Clinical Psychology, 49, 147-155.

Barlow, D. H., & Hayes, S. C., & Nelson, R. O. (1984). The scientist practitioner. New York: Pergamon.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill.

Boring, E. G. (1957). A History of experimental psychology (2nd ed.). New York: Appleton-Century-Crofts.

Botwin, M. D., & Buss, D. M. (1989). Structure of act-report data: Is the five-factor model of personality recaptured? Journal of Personality and Social Psychology, 56, 988-1001.

Brebner, M. T., & Welford, A. T. (1980). Introduction: An historical background sketch. In A. T. Welford (Ed.), Reaction times (pp. 1-24). London: Academic Press.

Bulkley, J. A. (1992). The prosecution's use of social science expert testimony in child sexual abuse cases: National trends and recommendations. Journal of Child Sexual Abuse, 1, 73-93.

Burisch, M. (1984). Approaches to personality inventory construction. American Psychologist, 39, 214-227.

Buros, O. K. (Ed.) (1970). Personality tests and reviews: I. Highland Park, NJ: Gryphon Press.

Calvert, E. J., & Waterfall, R. C. (1982). A comparison of conventional and automated administration of Raven's Standard Progressive Matrices. International Journal of Man-Machine Studies, 17, 305-310.

Canfield, M. L., Walker, W. R., & Brown, L. G. (1991). Contingency interaction analysis in psychotherapy. Journal of Consulting and Clinical Psychology, 59, 58-66.

Cattell, R. B. (1946). Description and measurement of personality. New York: World Book.

Cliff, N. (1992). Abstract measurement theory and the revolution the never happened. Psychological Science, 3, 186-190.

Cone, J. D. (1988). Psychometric considerations and the multiple models of behavioral assessment. In A. s. Bellack & M. Hersen (Eds.), Behavioral assessment: A prectical handbook (3rd ed., pp. 42-66). Elmsford, NY: Pergamon.

Cone, J. D., & Foster, S. L. (1991). Training in measurement: Always the bridesmaid. American Psychologist, 46, 653-654.

Coombs, C. H. (1964). A theory of data. New York: Wiley.

Cronback, L. J. (1946). Response sets and test validity. Educational and Psychological Measurement, 6, 475-494.

Cronbach, L. J. (1957). Beyond the two disciplines of scientific psychology. American Psychologist, 30, 116-127.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. American Psychologist, 30, 116-127.

Dahlstrom, W. G. (1985). The development of psychological testing. In G. A. Kimble & K. Schlesinger (Eds.) Topics in the history of psychology (Vol. 2, pp. 63-113. Hillsdale, HJ: Erlbaum.

Danzinger, K. (1990). Constructing the subject. Cambridge, England: Cambridge University Press.

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. American Psychologist, 42, 145-151.

Davis, C., & Cowles, M. (1989). Automatee psychological testing: Methods of administration, need for approval, and measures of anxiety. Educational and Psychological Measurement, 49, 311-320.

Davidson, M. L., Damarin, F., & Drasgow, F. (1986). Psychometrics and graduate education. American Psychologist, 41, 584-586.

Dawis, R. V. (1987). Scale Construction. Journal of Counseling Psychology, 34, 481-489.

Dawis, R.. V. (1992). The individual differences tradition in counseling psychology. Journal of Counseling Psychology, 39, 7-19.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. In M. R. Rosenzweig & L. W. Porter (Eds.), Annual review of Psychology (Vol. 41, pp. 417– 440). Palo Alto, CA: Annual Reviews.

Ellis, B. (1967). Measurement. In P. Edwards (Ed.), The encyclopedia of philosophy (Vol. 5, pp. 241-250). New York: Macmillan.

Embretson, S. E. (Ed.) (1985). Test design: Developments in psychology and psychometrics. New York: Academic Press.

Embretson, S. E., Schneider, L. M., & Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. Journal of Educational Measurement, 23, 13-32.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. Journal of Personality and Social Psychology, 37, 1097-1126.

Erdman, H. P., Klein, M. H., & Greist, J. H. (1985). Direct patient computer interviewing. Journal of Consulting and Clinical Psychology, 53, 760-773.

Exner, J. E., Jr. (1987). Computer assistance in Rorschach interpretation. In J. N. Butcher (Ed.), Computerized psychological asssessment (pp. 218-235). New York: Basic Books.

Falmagne, J. (1992). Measurement theory and the research psychologist. Psychological Science, 3, 88-93.

Faust, D., & Ziskin, J. (1988). The expert witness in psychology and psychiatry. Science, 241, 31-35.

Feldman, K. A. (1971). Using the work of others: Some observations on reviewing and integrating. Sociology of Education, 44, 86-102.

Forbes, R. J., & Dijksterhuis, E. J. (1963). A history of science and technology (Vol. 1). Baltimore, MD: Penguin Books.

Frederiksen, N. (1986). Toward a broader conception of human intelligence. American Psychologist, 41, 445-452.

Freeman, F. S. (1955). Theory and practice of psychological testing (rev. ed.). New York: Henry Holt.

Fretz, B. R., & Simon, N. P. (1992). Professional issues in counseling psychology: Continuity, change, and challenge. In S. D. Brown & R. W. Lent (Eds.), Handbook of counseling psychology (2nd ed.). Hew York: Wiley.

Friedman, E. H. & Sanders, G. G. (1992). Speech timing of mood disorders. Computers in Human Services, 8, 121-142.

Fulton, R. T., Larson, A. D., & Worthy, R. C. (1983). The use of microcomputertechnology in assessing and training communication skills of young hearing-impaired children. American Annals of the Deaf, 128, 570-576.

Gelso, C. J. (1979). Research in counseling: Methodological and professional issues. Counseling Psychologist, 8, 7-35.

George, M., & Skinner, H. (1990). Innovatie use of microcomputers for measureing the accuracy of assessment. In T. West, M. Christie, & J. Weinman (Eds.), Microcomputer, psychology, and medicine (pp. 251-266). Chichister, England: Wiley.

Gianetti, R. A. (1987). The GOLPH Psychosocial History: Response-contingent data acquisition and reporting. In J. N. Butcher (Ed.), Computerized psychological assessment (pp. 124-144). New York: Basic Books.

Gleick, J. (1987). Chaos. New York: Viking.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. Journal of Personality and Social Psychology, 59, 1216-1229.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. American Psychologist, 48, 26-34.

Goldman, L. (1990). Qualitative assessment. Counseling Psychologist, 18, 205-213.

Goldstein, G., & Hersen, M. (1990). Historical perspectives. In G. Goldstein & M. Herson (Eds.), Handbook of psychological assessment (2nd ed., pp. 3-17). New York: Pergamon.

Gould, S. J. (1981). The mismeasure of man. Norton: New York.

Graham, J. F. (1990). MMPI-2 Assessing personality and psychopathology. New York: Oxford University Press.

Green, C. D. (1992). Is unified positivism the answer to psychology's disunity? American Psychologist, 47, 1057-1058.

Gronlund, N. E. (1985). Measurement and evaluation in teaching (5th ed.). New York: Macmillan.

Guion, R. M. (1977). Content validity—The source of my discontent. Applied Psychological Measurement, 1, 1-10.

Gynther, M. D., & Green, S. B. (1982). Methodological problems in research with self-report inventories. In P.C. Kendall & J. N. butcher (eds.), Handbook of research methods in clinical psychology (pp. 355-386). New York: Wiley.

Hansen, J. C. (1987). Computer assisted interpretation of the Strong Interest Inventroy. In J. H. Butcher (Ed.), Computerized psychological assessment (pp. 292-324). New York: Basic Books.

Harrell, T. H., Honaker, L. M., Hetu, J., & Oberwager, J. (1987). Computerized versus traditional administration of the Multidimensional Aptitude Battery-Verbal scale: An examination of reliability and validity. Computers in Human Behavior, 3, 129-137.

Harrell, T. H., & Lombardo, T. A. (1984). Validation of an automated 16PF administrative procedure. Journal of Personality Assessment, 48, 638-642.

Hathaway, S. (1972). Where have we gone wrong? The mystery of the missing progress. In J. N. Butcher (Ed.), Objective personality assessment (pp. 24-44). New York: Academic Press.

Heidbreder, E. (1933). Seven psychologies. Englewood Cliffs, NJ: Prentice-Hall.

Hedl, J. J., O'Neil, H. F., & Hansen, D. N. (1973). Affective reactions toward computer-based intelligence testing. Journal of Consulting & Clinical Psychology, 40, 217-222.

Hedlund, J. L., Vieweg, B. V., & Cho, D. W. (1984). Mental health computing in the 1980s: II. Clinical applications, Computers in Human Services, 1, 1-31.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? American Psychologist, 47, 1083-1101.

Heppner, P. P., Kivlighan, D. M., & Wampold, B. R. (1992). Research design in counseling. Pacific Grove, CA: Brooks/Cole.

Herman, K. C. (1993). Reassessing predictors of therapist competence. Journal of Counseling and Development, 72, 29-32.

Hill, C. E. (1978). Development of a counselor verbal response category system. Journal of Counseling Psychology, 25, 461-468.

Hogan, R., & Nicholson, R. A. (1988). The meaning of personality test scores. American Psychologist, 43, 621-626.

Holden, C. (1988). Research psychologists break with APA. Science, 241, 1036.

Holden, R. R., Fekken, G. C., & Cotton, D. H. (1991). Assessing psychopathology using structured test-item response latencies. Psychological Assessment, 3, 111-118.

Honaker, L. M. (1988). The equivalency of computerized and conventional MMPI administration: A critical review. Clinical Psychology Review, 8, 561-577.

Howell, W. (1992). Field's science deficit will have dire results. APA Monitor, 23, 21.

Howell, W. (1993). Listen to academics—the future is talking. APA Monitor, 24, 22.

Jensen, A. R. (1982). Reaction time and psychometric g. In H. J. Eysenck (Ed.), A model for intelligence (pp. 93-132). New York: Springer-Verlag.

John, O. P., Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. European Journal of Personality, 2, 171-189.

Johnson, C. W., Hickson, J. F., Fetter, W. J., & Reichenbach, D. R. (1987). Microcomputer as teacher/researcher in a nontraditional setting. Computers in Human Behavior, 3, 61-70.

Judson, H. F. (1980). The search for solutions. New York: Holt, Rinehart & Winston.

Kaplan, A. (1964). The conduct of inquiry. San Francisco: Chandler, 1964.

Kazdin, A. E. (1980). Research design in clinical psychology. New York: Harper & Row.

Kendall, P. C., & Buckland, W. R. (1957). Dictionary of statistical terms. Edinburgh: Oliver & Boyd.

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

Lambert, N. M. (1991). The crisis in measurement literacy in psychology and education. Educational Psychologist, 26, 23-35.

Loevinger, J. (1957). Objective tests as instruments of psychological theory [Monograph Supplement No. 9]. Psychological Reports, 3, 635-694.

Loftus, E. F. (1993). The reality of repressed memories. American Psychologist, 48, 518-537.

Lohman, R. L. (1989). Human intelligence: An introduction to advances in theory and research. Research of Educational Research, 59, 333-373.

Lubin, B., Larsen, R. M., & Matarazzo, J. D. (1984). Patterns of psychological test usage in the United States: 1935-1982. American Psychologist, 39, 451-454.

Mahl, G. F. (1987). Explorations in nonverbal and verbal behavior. Hillsdale, NJ: Erlbaum.

Matarazzo, J. D. 1990). Psychological assessment versus psychological testing. American Psychologist, 45, 999-1017.

Matarazzo, J. D. (1991). Psychological assessment is reliable and valid: Reply to Ziskin and Faust. American Psychologist, 46, 882-884.

Matarazzo, J. D. (1992). Psychological testing and assessment in the 21st century. American Psychologist, 47, 1007-1018.

Mayer, T. P. (1966). Self—a measureless sea, St. Louis, MO: Catholic Hospital Association.

McCall, R. B. (1991). So many interactions, so little evidence. Why? In T. D. Wachs & R. Plomin (Eds.), Conceptualization and measurement of organism-environment interaction (pp. 142-161). Washington, DC: American Psychological Association.

McCrae, R. R., & Costa, P. T., Jr. (1985). Updating Norman's 'adequate taxonomy': Intelligence and personality dimensions in natural language and in questionnaires. Journal of Personality and Social Psychology, 49, 710-721.

McRae, R. R., & Costa, P. T., Jr. (1989). The structure of interpersonal traits: Wiggins' circumplex and the five-factor model. Journal of Personality and Social Psychology, 56, 586-595.

Meara, N. M., Shannon, J. W., & Pepinsky, H. B. (1979). Comparison of the stylistic complexity of the language of counselor and client across three theoretical orientations. Journal of Counseling Psychology, 26, 181-189.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.

Meehl, P. E. (1967). Theory testing in psychology and in physics: A methodological paradox. Science, 34, 103-115.

Meehl, P. E. (1991). Why summaries of research on psychological theories are often uninterpretable. In R. E. Snow & D. E. Wiley (eds.), Improving inquiry in social science (pp. 13-59). Hillside, NJ: Erlbaum.

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales use in counseling psychology research. Journal of Counseling Psychology, 37, 113-115.

Meier, S. T., & Wick, M. T. (1991). Computer-based unobtrusive measurement: Potential supplements to reactive self-reports. Professional Psychology: Research and Practice, 22, 410-412.

Mischel, W. (1968). Personality and assessment. New York: Wiley.

Mishler, E. G. (1986). Research interviewing. London: Harvard University Press.

Murphy, K. R., & Davidshofer, C. O. (1988). Psychological testing. Englewood Cliffs, NJ: Prentice Hall.

Murphy, J. T., Hollon, P. W., Zitzewitz, J. M., & Smoot, J. C. (1986). Physics. Columbus, OH: Charles E. Merrill.

National Institute of Mental Health. (1976). Putting knowledge to use: A distillation of the literature regarding transfer and change. Rockville, MD: Author.

Nettlebeck, T. (1973). Factors affecting reaction time: Mental retardation, brain damage, and other psychopathologies. In A. T. Welford (Ed.), Reaction times (pp. 355-402). London: Academic Press.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. Journal of Abnormal and Social Psychology, 66, 574-583.

Piotrowski, C., & Keller, J. W. (1984). Psychological testing: Trends in master's level counseling psychology programs. Teaching of Psychology, 11, 244-245.

Piotrowski, C., Keller, J. W. (1989). Psychological testing in outpatient mental health facilities: A national study. Professional Psychology: Research and Practice, 20, 423-425.

Piotrowski, C., & Lubin, B. (1990). Assessment practices of health psychologists: Survey of APA Division 38 clinicians. Professional Psychology Research and Practice, 21, 99-106.

Platt, J. R. (1977). Strong inference. In H. S. Broudy, R. H. Ennis, & L. I. Krimerman (Eds.), Philosophy of educational research (pp. 203-217). New York: Wiley.

Popham, W. J. (1993). Educational testing in America: What's right, what's wrong? Educational Measurement: Issues and Practice, 12, 11-14.

Reardon, R., & Loughead, T. (1988). A comparison of paper-and-pencil and computer versions of the Self-Directed Search. Journal of Counseling and Development, 67, 249-252.

Reckase, M. D. (1990). Scaling techniques. In G. Goldstein & M. Herson (Eds.), Handbook of psychological assessment (2nd ed., pp. 41-58). New York: Pergamon.

Rosenthal, R. (1976). Experimenter effects in behavioral research (rev. ed.). New York: Halsted Press.

Ryman, D. H., Naitoh, P., & Englund, C. E. (1984). Minicomputer-administered tasks in the study of effects of sustained work on human performance. Behavior Research Methods, Instruments and Computers, 16, 256-261.

Shertzer, B., & Stone, S. C. (1980). Fundamentals of counseling (3rd. ed.). Boston: Houghton Mifflin.

Shostrum, E. L. (Producer). (1966). Three approaches to psychotherapy. Santa Ana, CA: Psychological Films.

Sidman, M. (1960). Tactics of scientific research. New York: Basic Books.

Skinner, H. A., & Pakula, A. (1986). Challenge of computers in psychological assessment. Professional Psychology: Research and Practice, 17, 44-50.

Snow, R. E., & Wiley, D. E. (1991). Straight thinking. In R. E. Snow & D. E. Wiley (Eds.), Improving inquiry in social science (pp. 1-12). Hillsdale, NJ: Erlbaum.

Stein, S. J. (1987). Computer-assisted diagnosis for children and adolescents. In J. N. Butcher (Ed.), Computerized psychological assessment (pp. 145-158). New York: Basic Books.

Stone, E. F. (1978). Research methods in organizational behavior. Santa Monica, CA: Goodyear

Strong, E. K. Jr. (1943). Vocational interests of men and women. Stanford, CA: Stanford University Press.

Super, D. E. (1957). Psychology of careers. Harper: New York.

Tetrick, L. E. (1989). An exploratory investigation of response latency in computerized administrations of the Marlowe-Crowne Social Desirability Scale. Personality and Individual Differences, 10, 1281-1287.

Tinsley, D. J., & Irelan, T. M. (1989). Instruments used in college students affairs research: And analysis of the measurement base of a young profession. Journal of College Student Development, 30, 440-447.

Trapnell, P. D., & Wiggins, J. S. (1990). Extension of the Interpersonal Adjective Scales to include the Big Five dimensions of personality. Journal of Personality and Social Psychology, 59, 781-790.

Tryon, W. W. (Ed.). (1991). Activity measurement in psychology and medicine. New York: Plenum.

VanZandt, C. E. (1990) Professionalism: A matter of personal initiative. Journal of Counseling and Development, 68, 243-245.

Vansickle, T. R., Kimmel, C., & Kapes, J. T. (1989). Test-retest equivalency of the computer-based and paper-pencil versions of the Strong Campbell Interest Inventory. Measurement and Evaluation in Counseling and Development, 22, 88-83.

Wainer, H. (1993). Measurement problems. Journal of Educational Measurement, 30, 1-21.

Worthen, B. R., Borg, W. R., & White, K. R. (1993). Measurement and evaluation in the schools. New York: Longman.

Zimmer, J. M., & Cowles, K. H. (1972). Content analysis using FORTRAN: Applied to interviews conducted by C. Rogers, F. Perls, and A. Ellis. Journal of Counseling Psychology, 19, 161-166.

Ziskin, J., & Faust, D. (1991). Reply to Matarazzo, American Psychologist, 46, 881-882.