Some Emerging Trends in Psychological Measurement: A Fifty-Year Perspective

Anne Anastasi
Fordham University

Developments in psychological measurement over a 50-year period reveal a growing awareness of the modifiability of human behavior, as exemplified in cross-cultural comparisons, as well as in intergenerational changes within a single culture. Implications for testing are examined with special reference to the need for considering the context in which test takers developed and the contexts in which they are expected to function. Norms are viewed as the test performance of a population at a particular time and place. This orientation affects the interpretation of test scores. At a more basic level, cohort studies reveal systematic population changes. Cognitive scores may rise or decline, depending on concomitant societal changes. Progressive changes in attitudes, self-concepts, and other affective traits may in turn influence cognitive development. Affective variables may thus serve as intervening variables in the complex chain of events from genes to aptitudes. The traits identified by factor analysis are being increasingly recognized as descriptive categories for summarizing behavioral consistencies, rather than as underlying, fixed, causal entities. For testing purposes, this orientation provides flexibility in developing and choosing tests that fit the needed level, from highly specific behavioral units, through group factors of intermediate breadth, to such broad factors as scholastic aptitude. From a theoretical viewpoint, the question of factor formation becomes meaningful, insofar as the very traits into which intelligence becomes organized reflect the influence of individuals' learning histories and the experiential contexts in which they were reared.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 2, June 1985, pp. 121–138

0146-6216/85/020121-18$2.15

As I considered developments in psychological measurement from the 1930s to the 1980s, I identified a few major trends that impressed me as especially significant. For the purposes of this discussion, I have grouped them under three broad headings: (1) the contextual framework of testing, (2) norms and the longitudinal measurement of populations, and (3) the nature and formation of psychological traits. In each area, topics are approached from two viewpoints. One pertains to the applications of psychological measurement, insofar as there are implications for the construction and use of psychological tests and the interpretation of test scores. The other pertains to psychological theory and basic research, insofar as the results advance our understanding of the intrinsic characteristics and the etiology of the behavior assessed by tests.

The 50-year span surveyed in this paper includes its share of controversies, conflicting conclusions, and false starts. When viewed from a long-range perspective, however, these efforts have noticeably helped to mold the subsequent shape of the field, and their positive contributions are becoming integrated into the mainstream of current psychological measurement.

The Contextual Framework of Testing

Cross-Cultural Testing

From the earliest days of psychological testing, investigators began to use the available tests to

compare different cultural groups. The conclusions drawn from their findings were intrinsically ambiguous, because cultural differentiation and biological differentiation (or race formation) tend to be closely intertwined; both result chiefly from the isolation of human populations through geographical separation, customs, traditions, or any other conditions that reduce both interbreeding and cultural contact (Anastasi, 1958a, chap. 16). Although from the outset a few investigators recognized the difficulties of separating biological from cultural factors in accounting for the obtained group differences in test performance, the interpretations often reflected the investigator's theoretical orientation and the social climate of the period. Thus the early comparative studies, dating from the turn of the century through the 1930s and 1940s, identified the tested groups as races on the basis of their visible physical characteristics, and they assumed a biological basis for the performance differences. Later studies tended more and more to identify the groups in cultural terms and to look for explanations of group differences in the characteristic experiential history of their members.

Partly because of the available testing instruments and partly because of prevalent theories of racial differences, the earliest studies were concerned with sensory and perceptual differences. Among the first studies were those conducted at the St. Louis World's Fair in 1904 (F. G. Bruner, 1908; Woodworth, 1910; see also, Anastasi, 1958a, chap. 17). It was soon demonstrated, however, that the remarkable feats of certain aboriginal populations in the use of visual, auditory, or olfactory information to identify objects, interpret footprints, or find one's way resulted not from superior sensory equipment, but from learned responses to slight cues.

With the advent of intelligence tests, there was an influx of studies reporting mean differences in global test scores or IQs among many groups classified by race, nationality, geographical location, or other biological or cultural categories (Anastasi, 1958a, chap. 17; Garth, 1931; Klineberg, 1935; Mann, 1940). Among the earliest studies, one often finds lists of national and racial groups rank-ordered by mean scores on such global measures. With the development of separate tests for different aptitudes, it soon became apparent that the relative position of individual groups varied with the particular aptitude tested. A later and more significant methodological development was the joint investigation of group differences in test performance and in the culturally-based experiential or learning history of group members.

Cross-cultural investigators have been focusing increasingly on culture-specific experiences associated with typical differences in cognitive behavior (Cole & Bruner, 1971; Cole & Scribner, 1974; Rebelsky & Daniel, 1976; see also Anastasi, 1983a). A well-established finding is the difference among cultures in the extent to which behavior is linked to specific contexts. Some cultures provide relatively few opportunities for identifying common features across disparate experiences. Hence they do not encourage the generalization of learned cognitive skills. These differences are reflected in the development of abstract thinking and in the nature and breadth of the concepts that are formed.

There is evidence that formal schooling may play an important part in encouraging abstraction and concept formation. Research on preliterate cultures indicates that those members who have been exposed to a substantial amount of "Western-style" schooling are more likely to respond in terms of broad concepts and are less context-bound than their traditionally reared age peers (Cole, Gay, Glick, & Sharp, 1971; Cole & Scribner, 1974; Greenfield, 1966; Scribner, 1974; Sharp, Cole, & Lave, 1979). It should be noted that school learning typically occurs at specially designated times and places, a condition that serves to isolate the learner from the specific contexts of everyday life (J. S. Bruner, 1966; Scribner & Cole, 1973). In preliterate cultures, much of the traditional learning occurs through the child's performing real-life functions and participating in adult activities in their appropriate contexts. In the school setting, by contrast, functions and problems from many, disparate, real-life contexts are considered in a neutral setting, dissociated from their natural contexts. Knowledge is thus presented from a less restricted viewpoint and is evaluated in terms of universally applicable criteria.

The development of broadly generalizable concepts and logical analysis is further facilitated by certain technological advances, notably phonetic writing systems and numerical counting systems. Special attention has been given to the effects of a written language, in contrast to oral communication, as a means of codifying and transmitting the accumulated knowledge of a culture (Olson, 1976, 1977; Scribner & Cole, 1981). However, the mere availability of verbal and numerical symbol systems within a culture is likely to have little effect on the development of broad cognitive skills if these symbol systems are restricted to specific real-life contexts, such as letter writing or account keeping (Lancy, 1983; Reed & Lave, 1979; Scribner & Cole, 1978, 1981). It is the inclusion of these technologies in the relatively context-free setting of formal schooling that has a significant impact on the development of abstract thinking.

Culture-Free Tests

Although concern with cross-cultural testing has been greatly stimulated by recent social and political developments, the problem was recognized at least as early as 1910. Some of the earliest cross-cultural tests were designed for testing the large waves of immigrants coming to the United States at the turn of the century (Knox, 1914; see also, Anastasi, 1954a, chap. 10). Other early tests originated in basic research on the comparative abilities of relatively isolated cultural groups. Often, these groups had had little or no contact with Western civilization, within whose framework most psychological tests had been developed.

Typically, cross-cultural tests were designed to rule out one or more parameters along which cultures vary. Chief among such parameters are language, literacy (ability to read in *any* language), relative emphasis on the value of speed, and culture-specific test content. Early efforts to eliminate these culturally restricted features led to the design of the classic "culture-free" tests (Anastasi, 1954a, chap. 10; 1982, pp. 286–297). Examples include the Leiter International Performance Scale, Cattell's Culture-Free Intelligence Test, Raven's Progressive Matrices, and Goodenough's Draw-a-Man Test.

These early attempts to develop culture-free tests soon proved to be a false start, on both theoretical and empirical grounds. No test can be truly culture-free, because human behavior is not culture-free. Each culture requires and reinforces its own characteristic pattern of aptitudes. In practice, the so-called culture-free tests failed to come up to expectations (Anastasi, 1982, chap. 10). They were usually more heavily loaded with spatial and perceptual than with verbal and numerical aptitudes, and they did not prove as valid as the traditional intelligence tests in predicting educational and occupational criteria. Moreover, culturally diverse groups often performed no better on the culture-free tests than on other intelligence tests. Eventually, the term "culture-free" was dropped from test names, as well as from the general psychometric lexicon. Cattell's test, for example, is now titled "Culture Fair Intelligence Test." Some of the original test authors repudiated their earlier claims (e.g., Goodenough & Harris, 1950). Later revisions and uses of the original culture-free tests have been directed more and more toward other testing purposes. For example, the Goodenough-Harris Drawing Test (the current form of the Draw-a-Man Test) is oriented largely toward clinical applications.

Test Bias

In pluralistic societies such as the United States, the practical problems of cross-cultural testing have been associated chiefly with cultural minorities within the national culture. In this connection, the applicability of available tests to so-called culturally disadvantaged groups has often been questioned. It should be noted, however, that cultural disadvantage is a relative concept; objectively, there is only cultural difference. Each culture or subculture fosters and encourages the development of behavior that is adapted to its needs and values. When individuals must adjust to and compete within a cultural milieu other than that in which they were reared, cultural difference is likely to become cultural disadvantage.

It is in this context that the question of test bias has arisen (see Anastasi, 1982, pp. 183–191; 1985a, pp. 109–111). Essentially, test bias refers to group differences in test performance that are not reflected in corresponding differences in the behavior domain that the test is designed to assess. More specifically, test fairness (or absence of cultural bias) means that the test is equally valid for all groups with which it will be used and that it does not underpredict or overpredict the criterion performance of any group. In terms of the familiar regression model, these two goals refer to the avoidance of slope bias (validity differences) and intercept bias (underprediction or overprediction) of the regression lines, respectively (see Anastasi, 1982, pp. 182–191; 1985a, pp. 109–111; Cleary, 1968; Petersen & Novick, 1976). Test constructors have been giving special attention to ensuring that their tests meet these conditions. And there is a growing body of research with both educational and occupational criteria demonstrating that these conditions are satisfactorily met by most current ability tests.[1]

This approach differs from the earlier attempts to design culture-free or culture-fair tests. In the effort to include in such tests only functions common to many cultures, the test constructor was likely to choose content that had little relevance to any criterion domain to be assessed. A more effective solution is to choose criterion-relevant content to begin with and then investigate possible group differences in the regression of test scores on criterion measures.

Testing in Context

The testing of cultural minorities is only one aspect of a broader issue. Psychological measurement, like all psychology, needs to be fitted within a framework of cultural diversity. In order to formulate generalizations about human behavior, it is necessary to gather data and test hypotheses across different cultures (Azuma, 1984; Irving, Perl, Trickett, & Watts, 1984; Russell, 1984). For practical purposes, the most effective tests are likely to be those developed for clearly defined purposes and for use within specified contexts. Although these contexts will vary in breadth, none is likely to encompass the entire human species. The important point is to identify the locus and range of cultural (or other experiential) context for which any given test is appropriate. In choosing tests for a particular purpose and in interpreting test scores, the test user needs to be cognizant of the relevant cultural contexts.

We live in a pluralistic society, not only within large, heterogeneous nations such as the United States, but also within the broader, world-wide society. Since midcentury, there has been an increasing need for testing persons with highly dissimilar cultural backgrounds. This is illustrated by the use of tests for the maximum utilization of human resources in newly developing nations. The rapidly expanding educational facilities in these countries require testing for admission purposes as well as for individual counseling. With increasing industrialization, there is a mounting demand for tests to assist in the job selection and placement of personnel, particularly in mechanical, clerical, and professional areas.

All tests measure what the individual is able to do at the time. The interpretation of test scores, however, may involve either a backward or a forward reference. If we wish to understand why particular individuals perform as they do on a test, we need to look back into each individual's experiential background. If, on the other hand, we want to assess individuals' readiness to move into a course of study, a particular job, a career, a country to which they plan to immigrate, or an industrialized culture in a developing nation, we must ascertain whether they have acquired the prerequisite cognitive skills and knowledge for such a move. Tests used for these purposes should be based on the requirements of the *new* context into which individuals wish to advance and in which they will be

[1]In fact, in those studies finding a significant intercept bias, the criterion performance of minority groups is usually overpredicted, thereby favoring minority members in selection decisions based on the same cutoff score for all applicants (for theoretical rationale and empirical findings, see Anastasi, 1982, pp. 187–189; 1985a, pp. 109–111; Linn & Werts, 1971).

expected to function. In this sense, the tests have a forward, or predictive, reference.

From the standpoint of basic research, cross-cultural testing may be employed in investigating the contribution of one's reactional biography or learning history to the formation of psychological traits, the emergence of particular aptitudes, or the nature of human intelligence. Such research calls for indigenous tests, designed to assess socially significant behavioral constructs within each culture. In technologically advanced Western cultures, the constructs thus identified are illustrated by scholastic aptitude, verbal ability, and quantitative reasoning. Merely transplanting these constructs in studies of other cultures would fail to reveal important cognitive skills and knowledge fostered by different cultures.

To identify environmental demands in different contexts, there is need for indigenous task analyses and taxonomies of human performance (Fleishman, 1975; Fleishman & Quaintance, 1984; McCormick, Jeanneret, & Meacham, 1972,; Pearlman, 1980). Such research is required not only for cross-cultural investigations, but also for the study of specialized behavioral contexts within a highly developed and complex national culture. For example, with the increasing interest in life-span developmental psychology and the expansion of adult testing, there is a growing need for taxonomies of occupational behavior and other adult activities. An early effort to develop tests through task analyses of the everyday activities of older persons illustrated the differences in cognitive demands at different life stages (Demming & Pressey, 1957).

In summary, the systematic analysis of environmental demands should strengthen the basis for test construction. This approach would also encourage the development and use of tests for clearly specified purposes and within appropriate contexts, both within and across cultures.

Norms and the Longitudinal Measurement of Populations

Test norms should be regarded as a record of the performance of a specified population at a given time and place. This orientation goes beyond the periodic reevaluation and updating of norms. The more inclusive issue concerns population changes over time and their relation to societal changes. It is not only the test constructor but also the test user who must be alert to the conditions that affect norms. When interpreting the test scores of any given group or individual, the test user should take into account the specific influences that may have acted upon the normative population.

Suppose that an investigator finds that current engineering school applicants perform much better than the norms on a particular test. The investigator would undoubtedly consider the hypothesis that a more stringent self-selection (for a variety of possible reasons) may have made the current applicant sample superior in the tested abilities. It is equally likely, however, that a lower level of self-selection may have characterized the normative sample because of special societal pressures operating over a short atypical period. The point of this example is simply that test norms represent empirical data obtained at a particular time and place; they are subject to the same influences as is the test performance of any group. Despite the growing realization by test users that norms are not fixed and absolute, there is still a tendency to treat such norms as relatively immune to the conditions that affect the behavior of ordinary humans.

Societal Changes and the Interpretation of Test Scores

Revised editions and shifting IQs. Population changes in test performance may affect the interpretation of test scores in various ways. One of the best known examples is provided by the use of revised editions of intelligence tests. In both the restandardization of the Stanford-Binet (1937 vs. 1972)[2] and the revision of the Wechsler Intelligence Scale for Children (1949 vs. 1974), the later normative sample performed substantially better than

[2]The intervening 1960 revision of the Stanford-Binet, which introduced marked changes in item selection and placement, followed a norming procedure that related test performance to the 1937 standardization sample.

the earlier sample. As a result, the same children would receive lower IQs if tested with the revised edition than they would on the earlier edition, simply because their performance was evaluated against higher norms. The higher educational level of the parents of children in the later normative sample was one of the conditions cited for the rise in tested intelligence.

In adult testing, the influence of a rising educational level is reflected directly in the performance of the persons included in the normative samples for different editions. The implications for test interpretation were illustrated in a study with the Wechsler Adult Intelligence Scale (Wechsler, 1981). A sample of 72 persons between the ages of 35 and 44 who had taken both the 1955 and the 1981 editions obtained mean IQs of 111.3 on the earlier edition and 103.8 on the later edition. This 7.5-point difference suggests that the earlier norms were lower than those established more recently.

Aging and intellectual decline. Another potential misinterpretation of norms is illustrated by the use of adult age norms as a data base in the study of aging. Normative samples are chosen so that the persons at each age level are representative of the current population of that age. Hence, any other variables that differentiate among the age groups will be confounded with age, and their effects will be incorrectly attributed to aging (Anastasi, 1956a; 1958a, pp. 240–243; 1982, pp. 333–339).

In the Wechsler Adult Intelligence Scale, for example, the older members of the normative sample had completed fewer years of schooling, on the average, than had the younger members; accordingly, they scored lower on the test than did the younger members. As a result, intelligence appeared to decline steadily from the age of 30 on (Doppelt & Wallace, 1955). This score decline was widely accepted as an aging phenomenon and it is still so regarded by some professional practitioners. The findings corroborated a similar decline reported in early cross-sectional studies of aging, in which standardized tests were administered to adults of different ages at the same time period (Jones & Conrad, 1933; Miles & Miles, 1932; Wechsler, 1944). Later longitudinal studies, in which adults were retested after periods of 5 to 40 years, usually showed the opposite trend; the mean scores tended to improve with age, although the experiential history of individuals during the intervening years influenced the results (see Anastasi, 1982, pp. 335–336).

Neither cross-sectional nor longitudinal studies alone can provide a conclusive interpretation of observed age changes. On the one hand, age differences in educational level and other cultural advances may produce a spurious age decrement in test performance in cross-sectional studies. On the other hand, as individuals grow older, they are themselves exposed to cultural changes, such as expanding means of communication and transportation, that may improve their performance on intelligence tests. More recently, Schaie and his associates (Schaie, 1965; Schaie & Labouvie-Vief, 1974) have been following a composite *cross-sequential design,* which includes traditional cross-sectional and longitudinal study of individuals in combination with time-lag comparisons. The latter require the testing of same-age cohorts at different time periods. For instance, 20-year-olds tested in 1940 are compared with 20-year-olds tested in 1970. In general, these intercohort or intergenerational differences, which are associated with broad societal changes, account for much of the ability decrement formerly attributed to aging.

Population Changes in Intelligence Test Performance

The investigation of population changes over time may be characterized as the longitudinal study of populations (Anastasi, 1958a, pp. 209–211; 1962; 1982, pp. 339–341). The traditional application of the longitudinal method in psychology involves the repeated testing of the same individuals over time. In the longitudinal study of populations, on the other hand, the same population is sampled at different time periods. The comparison is between cohorts of persons born at different times but tested at the same ages.[3] Various procedures have been

[3]A special application of this general method can be recognized in the time-lag comparisons incorporated by Schaie (1965) in his previously cited cross-sequential design.

employed in these comparative studies. One procedure is to administer the identical test after a lapse of time, as was done in surveys of 11-year-old Scottish children in 1932 and 1947 (Scottish Council, 1949). Another procedure is to give two tests to a representative sample of persons in order to establish a correspondence between the two sets of scores and thereby "translate" performance from one test to the other. This was done in a comparison of the performance of soldiers in the U.S. Army in World Wars I and II, who had been examined with the Army Alpha and the Army General Classification Test, respectively (Tuddenham, 1948). A third, and technically sounder, approach is based on the establishment of an absolute, sample-free score scale through the use of anchor items, as is done with the College Board tests (Angoff, 1971; Donlon, 1984). The application of item response theory (IRT) techniques represents a further refinement of this approach.

Rising scores. Several large-scale American and British investigations conducted during the first five decades of the twentieth century revealed a rising intellectual level of the general population as measured by intelligence tests (see Anastasi, 1958a, pp. 209–211). These findings contradicted the intellectual decline predicted by several psychologists, geneticists, and demographers (e.g., Bradford, 1925; Burt, 1946; Cattell, 1937; Dawson, 1932). The anticipated decline was based on the widely reported negative correlation between intelligence and family size.

The discrepancy between the predicted decline and the actual rise in performance may be explained by at least two sets of conditions. First, research on the relation between intelligence and family size is beset with methodological pitfalls, especially with regard to sampling problems and statistical artifacts (Anastasi, 1954b, 1956b). Among the many uncontrolled and confounding variables are mortality rate, frequency of unmarried persons and childless couples, age of parents, and number of incomplete families in which the eventual number of offspring is still undetermined at the time of the investigation. Each of these variables is itself related to intellectual, educational, and socioeconomic levels. It is also noteworthy that, in most studies, the intelligence of children was correlated with the size of their sibship. A few studies in which parental intelligence was correlated with the number of offspring failed to find the usual negative correlation (see Anastasi, 1956b, pp. 194–195).

A second set of explanatory conditions is provided by the societal changes that paralleled the rising intelligence test scores, such as increasing literacy, higher educational levels, and expanded communication media. It appears that these cultural advances accounted in large part for the observed improvement in tested intelligence.[4] It is theoretically possible, of course, that adverse selective breeding caused a genuine intellectual decline that was masked and counterbalanced by the favorable societal changes. The predicted intellectual decline, however, is highly questionable, because of the methodological shortcomings of the studies that led to this prediction.

Declining scores. Whether the intelligence test scores of a given population rise, decline, or remain stable over time depends on many conditions. The time period covered, with its concomitant cultural changes, is clearly a prime factor. The age of the persons examined also makes a difference. For instance a rising educational level of the population will directly affect the test performance of adults, but it will only indirectly influence children's performance, since the children in successive cohorts will have had the same amount of education when tested. Another important consideration, especially when examining a selected subpopulation, is any change in the degree of selection occurring at different time periods. For example, if a larger proportion of the American population attended high school in 1960 than in 1910, as was actually the case, then the 1910 high school students represent a more highly selected sample of the general populations of their own time than do the 1960 high school students.

[4]Prior test-taking experience proved to be of minor importance in accounting for the observed improvement (Scottish Council, 1953, pp. 121–124; see also Anastasi, 1956b, p. 199).

The number and complexity of conditions that may account for a rise or decline in the tested intelligence of a population are clearly illustrated by an analysis of the highly publicized score decline on the College Board's Scholastic Aptitude Test (SAT). Between 1963 and 1977, the mean SAT Verbal score fell from 478 to 429, and the mean SAT Mathematical score fell from 502 to 470. In an effort to understand this steady 14-year score decline, a specially appointed panel commissioned 38 studies by experts in various areas and considered an impressive array of causal hypotheses (Wirtz, 1977).

A major conclusion reached by the panel was that the causal pattern differed from the first to the second half of the 14-year period. During the first seven years, the score decline resulted predominantly from a compositional change in the group taking the SAT. Because of the continuing increase in the number of high school graduates going to college over this period, the sampling became progressively less selected in the cognitive skills measured by the test. During the second seven years, however, the college-going population had become stabilized, and sampling selection accounted for a much smaller portion of the score decline. For this period, the explanation had to be sought in conditions in the home, the school, and society at large. The panel observed that the available data did not permit a determination of the relative contribution of different cultural changes to the score decline. Among the many factors cited as probably significant, however, were a diminished emphasis on academic standards, grade inflation and automatic promotions, reduced homework assignment, increased school absenteeism, diminished attention to mastery of skills and knowledge, excessive television viewing, and the many added distractions in students' lives arising from the social upheavals of the period. It should be added that the score decline, while most thoroughly investigated with reference to the SAT, was not limited to this instrument. Not only did it occur in other college admission tests, but there is also evidence of a corresponding decline in test performance at the high school and elementary school levels, during the same period.

Population Changes in Affective Variables

Most longitudinal cohort studies have been concerned with cognitive variables, especially as assessed by standardized intelligence tests. A few promising beginnings have been made, however, in studying population changes in affective characteristics. In this connection, *affect* is used in a broad sense, to cover all noncognitive characteristics, including feelings, moods, emotions, attitudes, interests, and motives.

An early exploratory study. A pioneer cohort study was conducted with the Pressey X-O Test (Pressey & Jones, 1955). The methodology and findings can be illustrated by considering a part of the study, in which participants were instructed to cross out all the things they considered wrong in a list of 125 items; most of the items referred to such borderline activities as smoking, drinking, spitting, and giggling. The test was administered in 1953 to college students and to general adult groups between the ages of 20 and 60 years. The same test had been given to comparable groups of college students in 1923, 1933, and 1943. The data thus permitted cross-sectional age comparisons within the 1953 sample and longitudinal cohort comparisons of college students tested over four decades.

The cross-sectional analysis revealed a clear tendency for the number of things considered wrong to increase with age in the general adult sample. Even at the youngest age level, moreover, these adults considered more things wrong than did the college students tested in the same year, a finding that suggests also an educational difference. The longitudinal population study of college students, on the other hand, showed strong cohort differences, with a sharp decrease in number of things considered wrong over the four decades. Of special interest is the comparison of adults tested in 1953 with students who were in college when these adults were in their twenties. For example, the 1953 50-year-olds marked about the same number of items as did the college seniors of 1923. Thus, it did not appear that persons tended to become more conservative and judgmental in their attitudes as they

grew older. Rather, the older persons tended to retain the attitudes characteristic of the cultural period in which they reached maturity. In fact the results suggested a slight movement in the direction of cultural change. Note that the adult scores were approximately equal to those of college students of their generation. In relation to a cohort with comparable education, the adult group would probably have scored lower.

Sex differences and cultural change. The study of sex differences in psychological characteristics is a particularly promising area for investigating the long-term effects of societal changes on behavior[5] (Anastasi, 1981). Stimulated by the feminist movement, research on sex differences assumed renewed vigor in the 1960s and began to advance in some new directions. Cohort studies of sex differences over critical periods of societal change should contribute to an understanding of the etiology of the observed behavioral differences. When data are limited to a relatively uniform and stable societal context, the results are likely to remain at a descriptive level. We can tell only how the behavior of females and that of males compare under the existing conditions. For obvious ethical and practical reasons, research on humans cannot expose participants to drastic and long-lasting variations in living conditions. In contrast, societal changes are likely to provide differences in experiential variables that are more extreme, of longer duration, and more pervasive in their influence on psychological development than could be achieved by experimental manipulation.

Cohort studies spanning one or more decades, during which significant social changes were under way, represent a valuable natural experiment—or at least a quasi-experiment. Even a simple comparison of findings obtained in the last quarter of the twentieth century with the published findings of earlier studies should contribute to an understanding of the origins of psychological sex differences, although systematic cohort studies with comparable samples and well-designed instruments would obviously yield more definitive answers. A beginning has been made in such cohort studies of sex differences in affective variables. The procedures vary, and the results are limited though suggestive. Examples include investigations of children's preferences for different play activities (Rosenberg & Sutton-Smith, 1960), achievement drive and so-called fear of success (Spence, 1974), and self-concepts and sex role conceptualizations (Urberg, 1979; Urberg & Labouvie-Vief, 1976).

Implications for cognitive development. Affective changes may in turn lead to cognitive changes. There is a growing body of data suggesting not only that transient affective states influence current performance, but also that more enduring personality traits influence the development of abilities (see Anastasi, 1982, pp 352–354; 1985b). Researchers investigating sex differences have been especially active in collecting evidence of the relationships between affective and cognitive variables. For instance, several studies show significant correlations between such affective variables as masculinity-femininity indices, sex role identification, and sex role conceptualizations, on the one hand, and the test performance of males and females in problem-solving (Milton, 1957), spatial aptitudes (Nash, 1979), and verbal and quantitative abilities (Dwyer, 1974; Entwisle & Baker, 1983; Fitzpatrick, 1978; Nash, 1979), on the other hand. In several of these studies, the relation between the attitudinal and cognitive variables was found not only when male and female groups were compared, but also *within* each sex.

It follows from such findings that, as sex role conceptualizations and expectations alter in response to societal events, sex differences in the development of certain aptitudes may also change. Cohort studies of cognitive functions, spanning critical periods of societal development, should reveal such changes in the relative performance of the sexes (Anastasi, 1981).

Methodological implications for the heredity-environment problem. The observed relations between affective and cognitive traits may also pro-

[5]Although most of the research specifically designed to analyze the effects of societal changes on affective and cognitive variables has been concerned with sex differences, surveys of available data on other subgroups have identified overall performance changes that parallel relevant societal changes (e.g., Jones, 1984).

vide a promising lead for investigating the contribution of hereditary and environmental conditions to individual differences. Insofar as attitudinal and motivational variables are shown to have a lasting influence on the development of abilities, they may represent an intermediate link in the chain of events from basic genetic and physiological factors to eventual individual differences in cognitive functioning.

Even before individual differences in aptitudes could be empirically assessed, the source of these differences in the individual's heredity and past environment was vigorously debated—and the debate continues. Some 25 years ago, I suggested that investigators may have been asking the wrong questions about heredity and environment and that a more fruitful approach might be to ask the question "How?" (Anastasi, 1958b). Rather than asking *which* differences are hereditary and which acquired, or *how much* of the variance is attributable to heredity and how much to environment, it would be better to investigate the modus operandi of hereditary and environmental influences in the development of individual differences. That genes could directly determine individual differences in, for example, verbal or mathematical aptitudes seems unlikely in the light of current knowledge about the development of human behavior. What is needed is more information about the many intervening steps in the etiological chain of events from genes to behavior. The role of motivation may represent one such step.

Although several investigators have acknowledged the contribution of motivation to intellectual development, its mediating role was most explicitly recognized by Hayes (1962), who wrote, "Intelligence is acquired by learning, and inherited motivational makeup influences the kind and amount of learning which occurs. The hereditary basis of intelligence consists of drives, rather than abilities as such" (p. 302). Hayes went on to discuss experience-producing drives and cited evidence for genetically controlled motivational differences between strains from several species, whether naturally occurring or produced by selective breeding.

A major methodological implication of the reformulated heredity-environment question is that attention is now focused on specific causal influences rather than on the broad, heterogeneous, and global domains of heredity and environment investigated in the traditional analysis of variance approach. It is more meaningful and more productive to use analysis of variance to assess the proportional contribution of more clearly defined variables within either domain (see also Lewontin, 1974; Mackenzie, 1984, especially pp. 1230–1232).

Nature and Formation of Psychological Traits

Identification and Organization of Factors

Current statistical techniques of factor analysis grew out of early investigations of the nature and composition of human intelligence. The first trait theory based on a statistical analysis of test scores was the two-factor theory proposed by the British psychologist, Charles Spearman (1904, 1914, 1927). In his original formulation, Spearman maintained that all intellectual activities share a single common factor, which he called the general factor, *g*. In addition, the theory postulated numerous specific or *s* factors, each strictly specific to a single test. Although the theory posits two types of factors, it is only the one *g* factor that accounts for correlation. In the terminology of current trait theories, therefore, it could be more precisely designated a single-factor theory, but the original label has survived.

Early in the development of his methodology, Spearman (1927) recognized that his two-factor theory needed to be qualified. When the tests are very similar, some correlation may be found over and above that attributable to *g*. Hence, there may be another type of factor, not so broad as *g* nor so strictly specific as the *s* factors. Such an intermediate factor, common to a group of intellectual activities but not to all, was designated a group factor. In his early writings, Spearman recognized the possible presence of very narrow and negligibly small group factors. This led to the practice of pooling tests before computing the correlations used in testing the two-factor theory. Following later investigations by colleagues and students, Spear-

man began to include broader group factors that corresponded to arithmetic, mechanical, and linguistic abilities, among others. Nevertheless, the major portion of common variance was still assigned to the *g* factor.

During the same period, trait research in the United States was focusing on multiple-trait theories. The publication of *Crossroads in the Mind of Man* by Truman L. Kelley (1928) paved the way for a number of studies in quest of particular group factors. The principal factors identified by Kelley in samples of schoolchildren were described as Verbal, Number, Memory, Spatial, and Speed factors. One of the leading exponents of multiple-factor theory was Thurstone. On the basis of extensive research, Thurstone proposed a small number of group factors, which he designated "primary mental abilities." Those most frequently corroborated in both his work and that of independent investigators include Verbal Comprehension, Word Fluency, Number, Space, Associative Memory, Perceptual Speed, and General Reasoning (Ekstrom, French, Harman, & Dermen, 1976; Thurstone, 1938; Thurstone & Thurstone, 1941). The next few decades saw a rapid proliferation of factors, in which several of Thurstone's factors were fractionated into narrower group factors. Such was the case, for example, when intensive investigations were conducted within areas originally identified as verbal, perceptual, memory, and reasoning.

One way of coping with the rapidly expanding number of factors is exemplified by Guilford's structure-of-intellect model (Guilford, 1967; Guilford & Hoepfner, 1971). The three dimensions of this model correspond to operations, content, and products; any test can be classified along all three dimensions (or facets). The model provides a total of 120 cells, in each of which at least one factor is expected, and some cells may contain more than one. The number of anticipated factors in this model is admittedly large, but Guilford argued that human nature is exceedingly complex and a few factors could not describe it adequately.

Another schema for classifying factors employs a hierarchical model, resembling an inverted tree (Burt, 1949; Humphreys, 1962; Vernon, 1961). At the top is a general factor; at the next level are broad group factors, similar to some of Thurstone's primary mental abilities; these major group factors subdivide into narrower group factors at one or more levels; the specific factors are at the bottom level. The hierarchical model permits the integration of Spearman's *g* with multiple-factor patterns. With regard to factor-analytic methodology, this reconciliation was made possible by the development of rotation techniques yielding oblique axes, which represent correlated factors. The intercorrelations among such factors can themselves be factor analyzed, thereby identifying second-order factors. With enough variables in the initial battery, still higher-order factors can be computed. Thurstone (1948) was among the first to point out that Spearman's *g* could emerge as a higher-order factor among correlated first-order factors. More recently, other factor analysts have been adopting hierarchical models, utilizing correlated factors and proceeding through several levels of higher-order factors (Cattell, 1963; Guilford, 1981; Hakstian & Cattell, 1978; Horn, 1968).

That no one level necessarily yields *the* basic or primary factors can be demonstrated through a method developed by Schmid and Leiman (1957) for transforming oblique factors of several orders into a single order of orthogonal factors defined by the original variables. It can thereby be shown that the difference between a first-order and a higher-order factor lies in the number of variables that define it; the higher-order factors are essentially broader factors. The investigator may choose that level of the hierarchy that is most appropriate for his or her purpose. This view of factors is gaining increasing recognition (Carroll & Horn, 1981; Coan, 1964; Humphreys, 1962).

A word of caution should be added regarding the general factor found in a particular battery. This factor has often been loosely described as Spearman's *g*. Actually, it represents a general factor common only to the tests in that battery. To conclude from such an analysis that a given test is heavily loaded with Spearman's *g* may therefore be misleading. It would be more meaningful to say that the general factor in that battery is heavily loaded with whatever that test measures, which can

be specified by examining the content of the test (e.g., verbal, mechanical). In fact, the best description of the general factor identified in a particular battery can be formulated by scrutinizing all the tests that have substantial loadings in that general factor. Thus, the unknown, hypothesized general factor would be explained in terms of the better known and closer-to-reality test scores from which it was derived. This is the procedure that is customarily followed in identifying other common factors.

Implications for Test Development and Use

From the standpoint of test construction, the hierarchical model combines a comprehensive theoretical framework with practical flexibility. The test constructor or test user may select that level of the hierarchy that is most suitable for a particular testing purpose. This solution corresponds to what is actually done in practice. Three levels of this hierarchy can be illustrated with types of tests in current use. At the most narrowly defined level are found some of the behavioral assessment techniques designed for use in programs of behavior modification and behavior therapy (see Anastasi, 1982, pp. 483–488 for summary and references). At about the same level of specificity are most so-called criterion-referenced tests,[6] especially those developed for use in individualized, self-paced instructional systems. In such educational programs, testing is closely integrated with instruction. It is introduced before, during, and after completion of each instructional unit in order to check on prerequisite skills, diagnose learning difficulties, and prescribe subsequent learning procedures (see Anastasi, 1982, pp. 94–101).

Multiple aptitude batteries focus on a much more broadly defined level of the trait hierarchy. Examples include the Primary Mental Abilities tests (PMA), which were a direct outgrowth of Thurstone's early factor-analytic research (Thurstone & Thurstone, 1946–1965), the Differential Aptitude Tests (DAT—Bennett, Seashore, & Wesman, 1947–1984), the General Aptitude Test Battery (GATB) prepared by the United States Employment Service (USES, 1946–1977), and several classification batteries developed by the military services.

The development and use of multiple aptitude batteries reached a peak in the 1950s, when it was widely believed that a few broad group factors could account for most of the variance of intellectual functioning. Hence, a profile of scores on tests measuring these factors should serve best in such classification problems as educational and vocational counseling and the assignment of personnel to occupational specialties that would maximally utilize each individual's pattern of cognitive skills and knowledge. Such applications presuppose high differential validity for the separate tests in the battery, in order to assess performance in different criterion situations. In this regard, however, the results proved disappointing. In the DAT, for example, it was soon found that the Verbal Reasoning test (VR) yielded high correlations with performance in most academic courses, regardless of content. It was chiefly because of such results that a composite score, representing the sum of scores on the Verbal Reasoning and Numerical Aptitude tests (VR + NA), was introduced as an index of scholastic aptitude.

A similar development is illustrated by ongoing research with the GATB. Reanalyses of data obtained for some 12,000 jobs listed in the *Dictionary of Occupational Titles* (U.S. Department of Labor, 1977) suggested that a large cognitive factor accounted for much of the variance of the battery and yielded high validity coefficients for most jobs. Additional measures of perceptual and psychomotor abilities improved predictive validity for some jobs, increasingly so for the less complex jobs (U.S. Department of Labor, 1983).

At the broadest level of the trait hierarchy are found the traditional intelligence tests. On the basis

[6]Because of the technical usage of the term criterion in psychometrics, *criterion-referenced* testing is something of a misnomer. In these tests, the interpretive frame of reference is a content domain rather than a population of persons. Hence, they can be more appropriately designated *content-referenced*. Other terms, such as *domain-referenced* and *objective-referenced*, have also been proposed, but criterion-referenced has survived as the most popular term.

of content, these tests closely resemble the combinations of tests that in the previously cited examples proved to be good predictors of both academic and occupational criteria. There is considerable accumulation of evidence showing that the particular cluster of cognitive skills and knowledge sampled by traditional intelligence tests plays a significant part in much of what goes on in modern, technologically advanced societies. Yet, the unqualified term *intelligence* is too broad to characterize such tests. There are many kinds of human intelligence: different cultural or experiential contexts foster and reinforce different sets of abilities, which constitute intelligence within those contexts. Current "intelligence" tests can be more precisely described as measures of academic intelligence or scholastic aptitude. They measure a kind of intelligent behavior that is both developed by formal schooling and prerequisite for progress within the academic system. Thus, in the interpretation of intelligence test scores, the concept of a segment of intelligence, albeit a broadly applicable and widely demanded segment, is replacing that of a general, universal human intelligence.

The Question of Situational Specificity

The concept of situational specificity has been the focus of controversy in more than one area of psychological measurement. Some confusion has also arisen because the term has been used in somewhat different senses in different contexts. In cognitive test research, it designates essentially specificity and diversity of situational demand, as illustrated by job requirements. In personality research on affective variables, it refers primarily to specificity and diversity of individual behavior in different situations.

With regard to the cognitive implications, it should be noted that tests are rarely, if ever used under conditions identical with those under which validity data were gathered. Hence, some degree of generalizability is inevitably required for practical test use. When standardized aptitude tests were first correlated with performance in presumably similar jobs in industrial validation studies, however, the validity coefficients were found to vary widely (Ghiselli, 1959, 1966). Such findings led to widespread pessimism regarding the generalizability of test validity across different situations. Until the mid-1970s, "situational specificity" of psychological requirements was generally regarded as a serious limitation in the usefulness of standardized tests in personnel selection.

More recently, research with newly developed statistical techniques has demonstrated that much of the variance among the validity coefficients reported for industrial samples may be a statistical artifact resulting from small sample size, criterion unreliability, and restriction of range in employee samples (Schmidt & Hunter, 1977). The subsequent accumulation of empirical evidence suggests that the validity of cognitive tests can be generalized far more widely across occupations than had heretofore been recognized (Ghiselli, 1973; Pearlman, Schmidt, & Hunter, 1980; Schmidt, Hunter, Pearlman, & Shane, 1979).

In the assessment of personality traits, on the other hand, situational specificity plays a more significant part (Anastasi, 1983b; Mischel, 1968, 1977; Mischel & Peake, 1982). For example, a person may be quite sociable at the office, but shy and reserved at social gatherings. An extensive body of empirical evidence has been assembled showing that individuals exhibit considerable situational specificity in several nonintellective dimensions, such as aggression, social conformity, dependency, and rigidity (see Anastasi, 1983b; Mischel, 1968; Peterson, 1968). Part of the explanation for the higher cross-situational consistency of cognitive than of affective functions may be found in the greater uniformity and standardization of the individual's reactional biography in the cognitive domain. Schooling is a major influence in the standardization of cognitive experience. The formal school curriculum, for example, fosters the development of broadly applicable cognitive skills in the verbal and numerical areas. Personality development, in contrast, occurs under far less uniform conditions. Moreover, in the personality domain, the same response may elicit social consequences that are positively reinforcing in one type of situation and negatively reinforcing in another. The

individual may thus learn to respond in quite different ways in different contexts.

The long-standing controversy between situational specificity and personality traits has been largely resolved (see Anastasi, 1983b). In the affective domain, traits may vary widely in breadth, as they do in the cognitive domain. Furthermore, some affective traits may need to be defined partly within situational boundaries. A well-known example is provided by test anxiety. Although this particular, situationally limited trait arose out of the concerns of psychometricians, personality theorists may likewise explore the usefulness of incorporating situational features in trait definitions.

Trait Formation

Factor analysis is no longer regarded as a means of searching for *the* basic, fixed, universal units of behavior. Rather, it is recognized as a method for organizing empirical data into useful categories through an analysis of behavioral uniformities. Like the test scores and other observational data from which they are derived, factors are descriptive, not explanatory. They do not represent underlying causal entities.

When the traits identified through factor analysis are viewed in this light, one can inquire into the causes of trait formation: what conditions lead to the development of particular behavioral consistencies or trait structures? As herein employed, the term *trait structure* should not be confused with performance profile, which shows the individual's relative standing in different traits—for example, whether a person is more advanced in verbal or in quantitative aptitude. Rather, trait structure refers to the organization of behavior into the very traits in terms of which the individual's performance is described. What brings about the correlational pattern that leads to the identification of a verbal trait and a quantitative trait in the first place? Several mechanisms have been proposed to account for trait formation (see Anastasi, 1983a, for references). One type of explanation focuses on the contiguity or co-occurrence of learning experiences; another is based on transfer of training, or the extent of generalizability of what is learned.

In the exploration of trait formation, some promising leads are provided by comparative studies of trait structures in populations with different experiential histories (Anastasi, 1970, 1983a). Let me cite one example from research on age differences. Factor-analytic investigations on elementary and high school populations, by both cross-sectional and longitudinal procedures, first led to the formulation of the *differentiation hypothesis* (Burt, 1954; Garrett, 1946). According to this hypothesis, intelligence is relatively undifferentiated in early childhood and becomes increasingly specialized throughout childhood and adolescence, with the emergence of distinct group factors in verbal, numerical, and spatial areas. Early exponents of the differentiation hypothesis, though recognizing the possible contributions of experiential variables, attributed age differentiation largely to maturation. However, fuller examination of the research, including some of the inconsistent findings of the earlier studies, suggests that the reported changes in factor pattern may be more closely associated with education than with age (Anastasi, 1948, 1970, 1983a; Khan, 1970, 1972).

In connection with such findings, one can speculate about the role that the structure of formal schooling may itself play in trait formation. As individuals advance through school, they encounter an academic curriculum that becomes increasingly differentiated into traditional subject-matter areas. Thus, instruction in verbal and numerical areas becomes gradually separated into different class periods and eventually is even given by different teachers. Similar separations occur with regard to instructional areas involving predominantly perceptual or spatial abilities, as in art, mechanical work, or other "practical" courses. This differentiation of academic experience is accompanied by an increasing prominence of narrower group factors. Moreover, there is evidence that different types of educational curricula, such as academic and technical, are associated with the development of qualitatively different group factors (Dockrell, 1965; Filella, 1960).

Suggestive evidence of the contribution of experiential history to trait formation has likewise been found in studies of groups representing dif-

ferent occupational, socioeconomic, and cultural contexts. There have also been a few attempts to investigate factor formation experimentally, either with animals or through short-term changes in humans. Essentially, what is needed is an open, dynamic orientation toward factor-analytic methodology, trait concepts, and the nature and composition of intelligence. Once we recognize that research findings in these areas are descriptive of human behavior at given times and places, we can proceed to devise ways of exploring the sources of the behavioral findings.

Postscript

In preparing this paper, I chose to discuss a few major developments in psychological measurement, observed over a 50-year span. Each of these developments was selected as intrinsically noteworthy, as were their specific aspects and manifestations. I did not undertake to present a unitary theme. As I look over the result, however, a common theme does emerge. It can be succinctly described as a growing recognition of the need to consider the environmental context of behavior in all psychological measurement.

This common theme, or general orientation, suggests certain directions for future research. It points toward the construction of tests with increasing reference to the contexts in which they will be used. And it calls for task analyses of behavior in a diversity of settings, as illustrated by occupational activities and other areas of significant adult behavior. These task analyses can be conducted within dissimilar cultures and across different life stages. The culturally significant behaviors thereby identified can then be measured and factor analyzed, as has been done in the traditional factor-analytic research with predominantly academic tasks and with student populations in Western-type schools.

The same theme is exemplified by seeking information about the experiential background and learning history of individual test takers, as well as of the normative samples on which tests are standardized. Such information should contribute to the effective interpretation of test scores for a myriad of purposes, ranging from the selection and placement of job applicants and the admission of candidates to educational programs, to individual counseling, the assessment of learning disabilities, and the identification of mentally retarded and intellectually talented children. It should also enhance our understanding of what makes people perform as they do and how their behaviors become organized into empirically identifiable trait categories.

References

Anastasi, A. (1948). The nature of psychological 'traits.' *Psychological Review, 55,* 127–138.

Anastasi, A. (1954a). *Psychological testing* (1st ed.). New York: Macmillan.

Anastasi, A. (1954b). Tested intelligence and family size: Methodological and interpretive problems. *Eugenics Quarterly, 1,* 155–160.

Anastasi, A. (1956a). Age changes in adult test performance. *Psychological Reports, 2,* 509.

Anastasi, A. (1956b). Intelligence and family size. *Psychological Bulletin, 53,* 187–209.

Anastasi, A. (1958a). *Differential psychology* (3rd ed.). New York: Macmillan.

Anastasi, A. (1958b). Heredity, environment, and the question "How?" *Psychological Review, 65,* 197–208.

Anastasi, A. (1962). The longitudinal study of populations. *Indian Psychological Bulletin, 7* (Part II), 25–28.

Anastasi, A. (1970). On the formation of psychological traits. *American Psychologist, 25,* 899–910.

Anastasi, A. (1981). Sex differences: Historical perspectives and methodological implications. *Developmental Review, 1,* 187–216.

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.

Anastasi, A. (1983a). Evolving trait concepts. *American Psychologist, 38,* 175–184.

Anastasi, A. (1983b). Traits, states, and situations: A comprehensive view. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A Festschrift for Frederic M. Lord* (pp. 345–356). Hillsdale NJ: Erlbaum.

Anastasi, A. (1985a). Psychological testing: Basic concepts and common misconceptions. In A. M. Rogers & C. J. Scheirer (Eds.), *The G. Stanley Hall Lecture Series* (Vol. 5, pp. 87–120). Washington DC: American Psychological Association.

Anastasi, A. (1985b). Reciprocal relations between cognitive and affective development—with implications for sex differences. In T. B. Sonderegger (Ed.), *Psychology and gender (Nebraska Symposium on Moti-*

vation, Vol. 32, pp. 1–35). Lincoln NB: University of Nebraska Press.

Angoff, W. H. (Ed.). (1971). *The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and achievement tests*. New York: College Entrance Examination Board.

Azuma, H. (1984). Psychology in a non-Western country, *International Journal of Psychology, 19,* 45–55.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1947–1984). *Differential Aptitude Tests*. Cleveland OH: The Psychological Corporation.

Bradford, E. J. G. (1925). Can present scholastic standards be maintained? *Forum of Education, 3,* 186–198.

Bruner, F. G. (1908). The hearing of primitive peoples. *Archives of Psychology,* No. 11.

Bruner, J. S. (1966). On cognitive growth II. In J. S. Bruner, R. R. Olver, & P. M. Greenfield (Eds.), *Studies in cognitive growth* (pp. 30–67). New York: Wiley.

Burt, C. (1946). *Intelligence and fertility*. London: Hamilton.

Burt, C. (1949). The structure of the mind; a review of the results of factor analysis. *British Journal of Educational Psychology, 19,* 100–111; 176–199.

Burt, C. (1954). The differentiation of intellectual ability. *British Journal of Educational Psychology, 24,* 76–90.

Carroll, J. B., & Horn, J. L. (1981). On the scientific basis of ability testing. *American Psychologist, 36,* 1012–1020.

Cattell, R. B. (1937). *The fight for our national intelligence*. London: King.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1–22.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Coan, R. W. (1964). Facts, factors, and artifacts: The quest for psychological meaning. *Psychological Review, 71,* 123–140.

Cole, M., & Bruner, J. S. (1971). Cultural differences and inferences about psychological processes. *American Psychologist, 26,* 867–876.

Cole, M., Gay, J., Glick, J. A., & Sharp, D. W. (1971). *The cultural context of learning and thinking*. New York: Basic Books.

Cole, M., & Scribner, S. (1974). *Culture and thought*. New York: Wiley.

Dawson, S. (1932). Intelligence and fertility. *British Journal of Psychology, 23,* 42–51.

Demming, J. A., & Pressey, S. L. (1957). Tests "indigenous" to the adult and older years. *Journal of Counseling Psychology, 4,* 144–148.

Dockrell, W. B. (1965). Cultural and educational influences on the differentiation of ability. *Proceedings of the 73rd Annual Convention of the American Psychological Association,* 317–318.

Donlon, T. F. (Ed.). (1984). *The technical handbook for the College Board Scholastic Aptitude Test and achievement tests*. New York: College Entrance Examination Board.

Doppelt, J. E., & Wallace, W. L. (1955). Standardization of the Wechsler Adult Intelligence Scale for older persons. *Journal of Abnormal and Social Psychology, 51,* 312–330.

Dwyer, C. A. (1974). Influence of children's sex role standards on reading and arithmetic achievement. *Journal of Educational Psychology, 66,* 811–816.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests, 3rd ed*. Princeton NJ: Educational Testing Service.

Entwisle, D. R., & Baker, D. P. (1983). Gender and young children's performance in arithmetic. *Developmental Psychology, 19,* 200–209.

Filella, J. F. (1960). Educational and sex differences in the organization of abilities in technical and academic students in Colombia, South America. *Genetic Psychology Monographs, 61,* 115–163.

Fitzpatrick, J. L. (1978). Academic underachievement, other-direction, and attitudes toward women's roles in bright adolescent females. *Journal of Educational Psychology, 70,* 645–650.

Fleishman, E. A. (1975). Toward a taxonomy of human performance. *American Psychologist, 30,* 1127–1149.

Fleishman, E. A., & Quaintance, H. K. (1984). *Taxonomies of human performance: The description of human tasks*. New York: Academic Press.

Garrett, H. E. (1946). A developmental theory of intelligence. *American Psychologist, 1,* 372–378.

Garth, T. R. (1931). *Race psychology: A study of racial mental differences*. New York: McGraw-Hill.

Ghiselli, E. E. (1959). The generalization of validity. *Personnel Psychology, 12,* 397–402.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests*. New York: Wiley.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26,* 461–477.

Goodenough, F. L., & Harris, D. B. (1950). Studies in the psychology of children's drawings: II. 1928–1949. *Psychological Bulletin, 47,* 369–433.

Greenfield, P. M. (1966). On culture and conservation. In J. S. Bruner, R. Olver, & P. M. Greenfield (Eds.), *Studies in cognitive growth* (pp. 225–256). New York: Wiley.

Guilford, J. P. (1967). *The nature of intelligence*. New York: McGraw-Hill.

Guilford, J. P. (1981). Higher-order structure-of-intellect abilities. *Multivariate Behavioral Research, 16,* 411–435.

Guilford, J. P., & Hoepfner, R. (1971). *The analysis of intelligence*. New York: McGraw-Hill.

Hakstian, A. R., & Cattell, R. B. (1978). Higher-stratum ability structures on a basis of twenty primary abilities. *Journal of Educational Psychology, 70,* 657–669.

Hayes, K. J. (1962). Genes, drives, and intellect. *Psychological Reports, 10,* 299–342.

Horn, J. L. (1968). Organization of abilities and the development of intelligence. *Psychological Review, 75,* 242–259.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17,* 475–483.

Irving, J., Perl, H., Trickett, E. J., & Watts, R. (1984). Minority curricula or a curriculum of cultural diversity? Differences that make a difference. *American Psychologist, 39,* 320–321.

Jones, H. E., & Conrad, H. S. (1933). The growth and decline of intelligence: A study of a homogeneous group between the ages of ten and sixty. *Genetic Psychology Monographs, 13,* 233–298.

Jones, L. V. (1984). White-black achievement differences: The narrowing gap. *American Psychologist, 39,* 1207–1213.

Kelley, T. L. (1928). *Crossroads in the mind of man: A study of differentiable mental abilities*. Stanford CA: Stanford University Press.

Khan, S. B. (1970). Development of mental abilities: An investigation of the "differentiation hypothesis." *Canadian Journal of Psychology, 24,* 199–205.

Khan, S. B. (1972). Learning and the development of mental ability. *Educational Research Journal, 9,* 607–614.

Klineberg, O. (1935). *Race differences*. New York: Harper.

Knox, H. A. (1914). A scale based on the work at Ellis Island for estimating mental defect. *Journal of the American Medical Association, 62,* 741–747.

Lancy, D. F. (1983). *Cross-cultural studies in cognition and mathematics*. New York: Academic Press.

Lewontin, R. C. (1974). The analysis of variance and the analysis of causes. *American Journal of Human Genetics, 26,* 400–411.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8,* 1–4.

Mackenzie, B. (1984). Explaining race differences in IQ: The logic, the methodology, and the evidence. *American Psychologist, 39,* 1214–1233.

Mann, C. W. (1940). Mental measurements in primitive communities. *Psychological Bulletin, 37,* 366–395.

McCormick, E. J., Jeanneret, P. R., & Meacham, R. C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology, 56,* 347–368.

Miles, C. C., & Miles, W. R. (1932). The correlation of intelligence scores and chronological age from early to late maturity. *American Journal of Psychology, 44,* 44–78.

Milton, G. A. (1957). The effects of sex-role identification upon problem-solving skills. *Journal of Abnormal and Social Psychology, 55,* 208–212.

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

Mischel, W. (1977). On the future of personality measurement. *American Psychologist, 32,* 246–254.

Mischel, W., & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89,* 730–755.

Nash, S. C. (1979). Sex role as a mediator of intellectual functioning. In M. A. Wittig & A. C. Petersen (Eds.), *Sex-related differences in cognitive functioning: Developmental issues* (pp. 263–302). New York: Academic Press.

Olson, D. R. (1976). Culture, technology, and intellect. In L. Resnick (Ed.), *The nature of intelligence* (pp. 189–202). Hillsdale NJ: Erlbaum.

Olson, D. R. (1977). From utterance to text: The bias of language in speech and writing. *Harvard Educational Review, 47,* 257–281.

Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. *Psychological Bulletin, 87,* 1–28.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65,* 373–406.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13,* 3–29.

Pressey, S. L., & Jones, A. W. (1955). 1923–1953 and 20–60 age changes in moral codes, attitudes, and interests, as shown by the "X-O Tests." *Journal of Psychology, 39,* 485–502.

Peterson, D. (1968). *The clinical study of social behavior*. New York: Appleton-Century-Crofts.

Rebelsky, F., & Daniel, P. A. (1976). Cross-cultural studies of infant intelligence. In M. Lewis (Ed.), *Origins of intelligence: Infancy and early childhood* (pp. 279–297). New York: Plenum.

Reed, H. J., & Lave, J. (1979). Arithmetic as a tool for investigating relations between culture and cognition. *American Ethnologist, 6,* 568–582.

Rosenberg, B. G., & Sutton-Smith, B. (1960). A revised conception of masculine-feminine differences in play activities. *Journal of Genetic Psychology, 96,* 165–170.

Russell, R. W. (1984). Psychology in its world context. *American Psychologist, 39,* 1017–1025.

Schaie, K. W. (1965). A general model for the study of developmental problems. *Psychological Bulletin, 64,* 92–107.

Schaie, K. W., & Labouvie-Vief, G. (1974). Generational versus ontogenetic components of change in cognitive behavior: A fourteen-year cross-sequential study. *Developmental Psychology, 10,* 305–320.

Schmid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53–61.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization model. *Personnel Psychology, 32,* 257–281.

Scottish Council for Research in Education. (1949). *The trend of Scottish intelligence*. London: University of London Press.

Scottish Council for Research in Education. (1953). *Social implications of the 1947 Scottish mental survey*. London: University of London Press.

Scribner, S. (1974). Developmental aspects of categorizing recall in West African society. *Cognitive Psychology, 6,* 475–494.

Scribner, S., & Cole, M. (1973). The cognitive consequences of formal and informal education. *Science, 182,* 553–559.

Scribner, S., & Cole, M. (1978). Literacy without schooling: Testing for intellectual effects. *Harvard Educational Review, 48,* 448–461.

Scribner, S., & Cole, M. (1981). *The psychology of literacy*. Cambridge MA: Harvard University Press.

Sharp, D., Cole, M., & Lave, C. (1979). Education and cognitive development: The evidence from experimental research. *Monographs of the Society for Research in Child Development, 44,* (1, 2, Serial No. 178).

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15,* 201–293.

Spearman, C. (1914). The theory of two factors. *Psychological Review, 21,* 101–115.

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

Spence, J. T. (1974). The Thematic Apperception Test and attitudes toward achievement in women: A new look at the motive to avoid success and a new method of measurement. *Journal of Consulting and Clinical Psychology, 42,* 427–437.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs,* No. 1.

Thurstone, L. L. (1948). Psychological implications of factor analysis. *American Psychologist, 3,* 402–408.

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs,* No. 2.

Thurstone, L. L., & Thurstone, T. G. (1946–1965). *SRA Primary Mental Abilities*. Chicago: Science Research Associates.

Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. *American Psychologist, 3,* 54–56.

United States Department of Labor, Employment and Training Administration. (1977). *Dictionary of Occupational Titles* (4th ed.). Washington DC: U.S. Government Printing Office.

United States Department of Labor, Employment and Training Administration. (1983). *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery* (USES Test Research Report No. 45). Washington DC: U.S. Government Printing Office.

United States Employment Service. (1946–1977). *USES General Aptitude Test Battery*. Washington DC: U.S. Government Printing Office.

Urberg, K. A. (1979). Sex role conceptualization in adolescents and adults. *Developmental Psychology, 15,* 90–92.

Urberg, K. A., & Labouvie-Vief, G. (1976). Conceptualizations of sex roles: A life span developmental study. *Developmental Psychology, 12,* 15–23.

Vernon, P. E. (1961). *The structure of human abilities* (Rev. ed.). London: Methuen.

Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). Baltimore: Williams & Wilkins.

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale—Revised*. New York: Psychological Corporation.

Wirtz, W. (Chair). (1977). *On further examination: Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline*. New York: College Entrance Examination Board.

Woodworth, R. S. (1910). Race differences in mental traits. *Science, 31,* 171–186.

Author's Address

Send requests for further information to Anne Anastasi, Department of Psychology, Fordham University, Bronx NY 10458, U.S.A.

Reprints

Reprints of this article may be purchased *prepaid* for $2.50 for delivery in the U.S. or $3.00 (in U.S. funds drawn on a U.S. bank) elsewhere, from Applied Psychological Measurement, N658 Elliott Hall, University of Minnesota, Minneapolis MN 55455, U.S.A.