DOCUMENT RESUME

ED 150 189 TM 006 923

TITLE	Assessment in a Pluralistic Society.
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	73
NOTE	139p.; Proceedings of the ETS Invitational Conference on Testing Problems (33rd, New York, New York, October 28, 1972)
AVAILABLE FROM	Invitational Conference, Educational Testing Service, Princeton, New Jersey 08541 ($5.00)
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	Academic Ability; Achievement Tests; Awards; Biological Influences; Black Colleges; Black Studies; Cognitive Development; Compensatory Education Programs; *Conference Reports; Cultural Differences; *Cultural Pluralism; Disadvantaged Youth; Early Childhood Education; *Educational Assessment; Elementary Secondary Education; Evaluation Methods; Group Norms; Higher Education; *Intelligence Differences; Mathematics; *Nature Nurture Controversy; Racial Differences; Sex Differences; Social Differences; Social Influences; Test Bias; *Testing Problems; Testing Programs; Test Interpretation; Test Reliability; Test Validity; Verbal Ability

ABSTRACT

Robert L. Thorndike was awarded the Educational Testing Service (ETS) Measurement Award at the 1972 ETS Invitational Conference. In "Heredity, Environment, and Class or Ethnic Differences," J. McV. Hunt addressed several fundamental questions pertaining to the hereditary and environmental influences of the observed social class and ethnic differences on intelligence and scholastic achievement tests. Eleanor E. Maccoby and Carol Nagy Jacklin, in "Sex Differences in Intellectual Functioning," summarized the current state of knowledge regarding sex differneces in verbal, mathematical, and spatial abilities. In his paper on "Implications of Group Differences for Test Interpretation," Lloyd G. Humphreys examined the concept of test fairness. Charles V. Willie discussed the previous papers in "A Theoretical Approach to Cultural and Biological Differences." In "Recycling the Problems in Testing," Henry S. Dyer observed that much of the misuse of tests arises from the diverse perceptions held by makers, givers, takers, and users of tests. The afternoon session focused on educational implementation: Irving E. Sigel spoke on "Where is Preschool Education Going: Or Are We En Route Without a Road Map?," and Benjamin F. Payton discussed "Black Colleges and Black Studies." (BW)

ASSESSMENT IN A PLURALISTIC SOCIETY

ASSESSMENT IN A PLURALISTIC SOCIETY

PROCEEDINGS OF THE
1972 INVITATIONAL
CONFERENCE
ON TESTING PROBLEMS

ETS
EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY

ATLANTA GEORGIA
AUSTIN TEXAS
BERKELEY CALIFORNIA
DURHAM NORTH CAROLINA
EVANSTON ILLINOIS
LOS ANGELES CALIFORNIA
WASHINGTON D C
SAN JUAN PUERTO RICO

The thirty-third Invitational Conference on Testing Problems, sponsored by Educational Testing Service, was held at the New York Hilton, New York City, on October 28, 1972.

Chairman: Anne Anastasi
Fordham University

PHOTO CREDITS: page x, Tommy Weber, New York
page xvi, Blackstone-Shelburne, New York
page 37, courtesy of Stanford University
page 85, Orren Jack Turner, Hopewell, New Jersey
page 99, The Merrill-Palmer Institute

Library of Congress Catalog Number: 47-11226
Printed in the United States of America

Contents

ETS Board of Trustees, 1972-1973

*George F. Baughman, *Chairman of the Board*
President
The College Foundation

*John T. Caldwell
Chancellor
North Carolina State University

Arland F. Christ-Janer
President
College Entrance Examination Board

John B. Davis Jr.
Superintendent of Schools
Minneapolis, Minnesota

Joseph W. Fordyce
President
Junior College District of St. Louis

*Hugh M. Gloster
President
Morehouse College

Kurt M. Hertzfeld
Treasurer
Amherst College

*Roger W. Heyns
President
American Council on Education

Wayne H. Holtzman
President, Hogg Foundation for Mental Health and Hogg Professor of Psychology and Education
The University of Texas at Austin

Juanita M. Kreps
James B. Duke Professor of Economics
Duke University

Malcolm C. Moos
President
University of Minnesota

*John A. Perkins
Vice President—Administration
University of California

Paul C. Reinert, S.J.
President
St. Louis University

Wilson C. Riles
Superintendent of Public Instruction
State of California

*William W. Turnbull
President
Educational Testing Service

Charles E. Young
Chancellor
University of California at Los Angeles

*Member of Executive Committee

ETS Officers

President

William W. Turnbull

Executive Vice President

Robert J. Solomon

Vice Presidents

Samuel B. Gould
John S. Helmick
Richard S. Levine
Winton H. Manning
Samuel J. Messick
Charles E. Scholl
John Summerskill
E. Belvin Williams

Vice President and Treasurer

David J. Brodsky

Assistant Vice President

Bernard L. Tchorni

Secretary

Catherine G. Sharp

Assistant Treasurers

Russell W. Martin Jr.
Herman F. Smith

Foreword

The pluralistic nature of our society poses a formidable challenge to those who attempt to measure human capabilities and evaluate human achievements. As Chairman of the 1972 Invitational Conference, Anne Anastasi faced this challenge squarely. She succeeded in assembling a group of speakers who are eminently qualified to speak authoritatively on the basic issues related to the measurement and evaluation problems of a pluralistic society. The large, attentive audience and the searching questions addressed to the speakers attest to her talent for delineating issues and finding the right people to interpret them.

Each of the speakers in his or her own way conveyed a sense of both continuity and urgency. Our problems are not new, and much of the relevant information has been available for many years; yet they resist solution. All of the speakers, and particularly Henry Dyer, who chose the recurrence of problems as a central theme for his luncheon speech, remind us that despite their intransigence these problems deserve the best effort we can muster.

Last year I expressed my hope that this decade would see a new and much-needed synthesis of education and measurement. I believe that the 1972 Conference represents a significant step in that direction.

William W. Turnbull
PRESIDENT

Anne Anastasi

Preface

The general theme of the 1972 Invitational Conference on Testing Problems was *Assessment in a Pluralistic Society*. This is a broad theme that could be considered from the viewpoint of many disciplines and could concentrate on many different concerns and problems. The approach followed in this Conference was to inquire into the contributions that basic research in psychology and related sciences can make toward the improvement of assessment. The underlying question was: How can our present knowledge about the nature and origins of individual and group differences in behavior improve the development and use of assessment procedures and the interpretation of assessment results? The term "assessment" was chosen advisedly, to cover not only testing but also other procedures for observing and evaluating human behavior. Although much of what the speakers reported was expressed with reference to tests, their conclusions and the implications of their findings generally apply equally well to other observational techniques, such as interviews, ratings, and records of job or school performance.

Psychometrics has made rapid progress in the refinement of techniques of test construction and the statistical evaluation of test results. At the same time, the increasing specialization of both scholarly and professional activities has tended to dissociate psychological testing from the mainstream of contemporary psychology. As psychometricians concentrate more and more on the testing process, they may lose sight of the behavior they set out to measure. As a result, outmoded interpretations of test scores may remain insulated from the findings of subsequent behavioral research. The widespread misconceptions about the so-called IQ provide a particularly flagrant example of such a dissociation. One still hears the term "IQ" used as though it referred, not to a test score, but to a property of the organism. It was a major objective of the 1972 Invitational Conference to encourage a rapprochement between testing and behavioral research and to illustrate some points of contact between the two fields.

In the morning session the focus was on understanding: What do we know about human behavior and what does such knowledge imply

for assessment? The three papers provided examples of the type of answers that contemporary psychology can give to these questions. In the opening paper, J. McV. Hunt addressed himself to several fundamental questions pertaining to the operation of hereditary and environmental factors in the development of observed social-class and ethnic differences on intelligence and scholastic achievement tests. The discussion was illustrated with research on the influence of early experience on the performance of animals and human infants. Such concepts as heritability and IQ constancy were analyzed.

The second speaker, Eleanor Maccoby, summarized the current state of knowledge regarding sex differences in verbal, mathematical, and spatial abilities, with particular attention to developmental changes from infancy to adolescence. The age-old question of sex differences in variability was reexamined in the light of recent data. Finally, in a discussion of the biological and experiential origins of intellectual sex differences, Dr. Maccoby critically reviewed several etiological hypotheses and pointed to the need for more solid data to test such hypotheses.

In his paper on "Implications of Group Differences for Test Interpretation," Lloyd Humphreys examined the concept of "test bias" or "test fairness," with specific reference to the accuracy of inferences drawn from test scores for males and females, ethnic minorities, and socioeconomic and regional subgroups. Various proposed solutions for reducing test bias in selection were compared and critically evaluated. Dr. Humphreys emphasized the practical value of tests in identifying existing learning deficiencies, whatever their cause, and observed that unfairness results not from the use of tests but from the incorrect interpretation of test scores as measures of learning potential and of fixed, innate capacities.

In order to broaden the perspective and provide a fresh look at the topic from a different angle, a representative of a related field was chosen as discussant for the three preceding papers. Charles Willie, Professor of Sociology and Vice President for Student Affairs at Syracuse University, was invited to perform this challenging task. Dr. Willie cited evidence reported by each of the speakers in support of the plasticity of human psychological development and its susceptibility to environmental influences. He also referred to the biological value of genetic polymorphism within a human population—a value that parallels the cultural advantages of a pluralistic society. His own research on the performance of black students in predominantly

white colleges demonstrated the overlapping ranges of intellectual capacity found among blacks and whites.

In the luncheon address, Henry Dyer pointed out that psychometricians and educators have for many years expressed concern about the misuse of tests. Today, however, the problem has assumed new magnitude, involving potential misinterpretation of test results not only by school personnel but also by politicians and their pluralistic constituencies. Observing that the misuse of tests arises largely from the diverse perceptions of testing held by the makers, givers, takers, and users of tests, he illustrated his point with regard to two of the six possible pairs among these four classes: makers-users and givers-takers. To the previously cited dissociation between psychometrics and behavioral research, Dr Dyer thus added other dimensions of compartmentalization, the practical consequences of which may be even more serious and far-reaching.

In the afternoon session, the focus was on educational implementation. Assessment is not usually an end in itself, but rather a means to some other, practical end. The development of appropriate educational programs is one major objective of assessment. Accordingly, the two afternoon speakers considered educational programs designed to meet the diverse needs of a pluralistic culture. The specific programs discussed were drawn from widely separated levels of the educational ladder: from the preschool in one case and from college in the other.

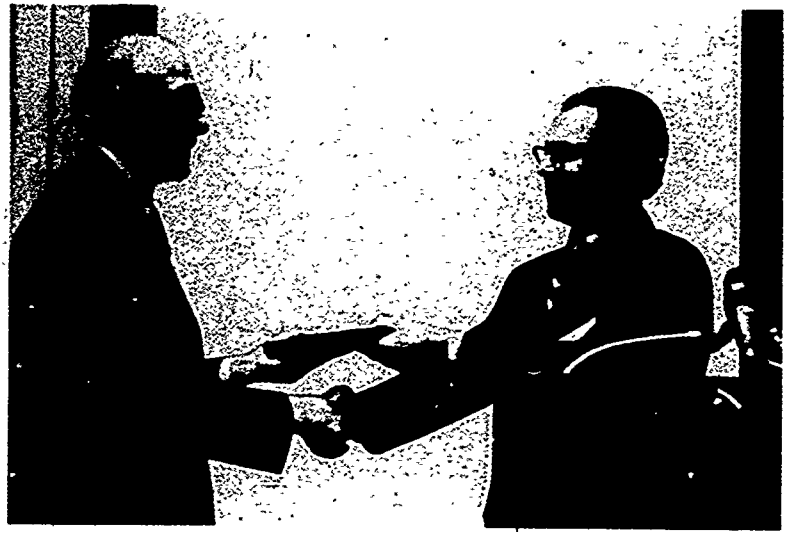
Irving Sigel called attention to the resurgence of interest in early childhood education in the 1960s—an interest too often manifested, however, in poorly planned crash programs that led to negative results and mounting disillusionment. In contrast, a greatly expanded knowledge base from psychology and related disciplines is now available. Dr. Sigel argued convincingly for three innovations: sophisticated reconceptualization of developmental processes, the use of specific achievement tests to measure progress in the skills developed by educational programs (instead of such global scores as "IQs"), and a reformulation of the goals of early childhood education in more realistic terms. In the appendix to his talk, these points are illustrated by reference to his own ongoing research in the Early Childhood Education Project at the State University of New York at Buffalo.

In the closing paper of the Conference, Benjamin Payton presented black colleges and black studies programs as means of preserving cultural pluralism. He noted that, historically, black colleges have been primarily responsible for the educational advancement of the

black population in America; and he argued that this type of institution still has an important role in current efforts to achieve equality of status for minority groups.

For the success of this Invitational Conference, primary credit is due to the speakers and invited discussant for providing illuminating and potent distillates of their many years of research and thought on the topics they covered. Thanks are also extended to the distinguished audience for its active interest and support, and especially to those who contributed searching and provocative questions from the floor. Some of these questions, together with the speakers' replies, are reproduced. From the Conference Chairman's special observation post, it is abundantly clear that the professional quality and smooth operation of the Conference are attributable to the imaginative, competent, and tireless activity of many members of the ETS staff. To name individuals would produce a very long list indeed, and even then some would be missed who had toiled mightily behind the scenes to bring off once more what has become a major annual event in psychometric circles. At the risk of sounding like the traditional acknowledgments in the front pages of doctoral dissertations, however, I feel impelled to record a special indebtedness to Anna Dragositz, without whose expertise, judgment, and good humor in the face of the inevitable frustrations and delays this Conference could not have taken place. My numerous contacts with her were invariably productive and a source of genuine personal pleasure for me.

Anne Anastasi
CHAIRMAN



Robert L. Thorndike (left) receiving the 1972 ETS *Measurement Award from* ETS *President Turnbull at the Conference luncheon.*

One president to another: Mr. Turnbull congratulates EDUCOM *President Henry Chauncey, Chairman-elect of the 1973 Invitational Conference on Testing Problems.*

Robert L. Thorndike

EDUCATIONAL TESTING SERVICE

Measurement Award

1972

The ETS Award for Distinguished Service to Measurement was established in 1970, to be presented annually to an individual whose work and career have had a major impact on developments in educational and psychological measurement. The 1972 Award was presented at the Conference by ETS President William W. Turnbull to Professor Robert Ladd Thorndike with the following citation:

A need to get to the heart of the matter and the intelligence to make it possible have been the foundation of Robert L. Thorndike's distinguished career in education and measurement.

He has provided clear and logical formulations in a number of important and complex areas, focusing on such concepts as the reliability of tests, the meaning and measurement of underachievement, and the assessment of test bias. A capacity for powerful analysis and precise writing has contributed to the effectiveness of his leadership in the collaborative revision of a major summative work, The American Council on Education's *Educational Measurement*.

As a scholar, he has combined the theory of measurement with its practice in developing a variety of instruments for assessing psychological traits and educational attainments. Further, through his gifts as a teacher and author, he has given thousands of students the skills to create and interpret instruments of their own.

Robert Thorndike's research studies are distinguished by their elegance and by the importance of the questions they address. Our understanding of the determinants of career choice and the correlates of career success, for example, has been furthered by his research. And we are indebted to him for his studies of such diverse areas as human problem solving and the comparative outcomes of education in different nations.

For his many contributions to the theory and practice of educational measurement, ETS is pleased to present the 1972 Award for Distinguished Service to Measurement to Robert L. Thorndike.

Session I

Origins and Implications of Differences

Heredity, Environment, and Class or Ethnic Differences[1]

J. McV. HUNT
University of Illinois
Champaign

That people differ on the average according to the social classes and the ethnic or racial groups to which they belong is an ancient observation. Within our own United States, the mean values or the incidence of behavioral phenomena have been observed to vary significantly for nearly every characteristic where systematic measurement has been tried. Such characteristics include emotional adjustment as measured by both the Minnesota Multiphasic Personality Inventory (Gough, 1946, 1948) and the Vineland Social Maturity Scale (Sims, 1954); the incidence of behavior disorders (Clark, 1949; Faris & Dunham, 1939; Landis & Page, 1938); the incidence of delinquent and criminal behavior (Shaw, 1929; Short & Strodtbeck, 1965); persistence in goal striving (Battle & Rotter, 1963; Hertzig, et al., 1968; Zigler & Butterfield, 1968); the tendency to be impulsive rather than reflective (Meichenbaum & Goodman, 1969; Mumbauer & Miller, 1970); measures of ability; and measures of both scholastic and occupational achievement. Such differences serve to define the social-class structure as described by Warner and his colleagues (see Warner, et al., 1949). Differences among ethnic and racial groups include a similar variety of characteristics plus differences of language, skin color, and cultural attitude.

[1]The preparation of this paper was supported by grants from the United States Public Health Service: MH-K6-18567 and MH-11321.

Especially prominent have been the social-class, ethnic, and racial differences in performance on tests of intelligence and scholastic achievement. Since the evidence is most abundant for such measures, they will be the focus of this paper. Where social class is the focus, the IQs of children entering school from professional families average about 115 while those from unskilled laborers average about 95 (McNemar, 1942), and differences in mean IQ of this order of 20 points are typical of those reported by other investigators (see Anastasi, 1958, Ch. 15 on Social Class Differences). Where race is the focus, black children usually average an IQ about one standard deviation, or 15 points below white children (Shuey, 1966). Moreover, Mayeske (1971) has described a special analysis of the data in the report on equality of educational opportunity by Coleman and others (1966) which shows that 25 percent of the total variance among American students in their academic achievement was associated with membership in one of six ethnic-racial groups: Indian, Mexican, native white, Negro, Oriental, and Puerto Rican.

Although some individuals have risen out of poverty to top levels of excellence, there can be no blinking these class and race differences. They exist. But are they biologically inevitable to the degree in which they now manifest themselves? Are class differences an inevitable matter of competitive social selection which has resulted from genotypic limits on potential? Or, might one expect most of such differences as exist from the conditions of child-rearing in the various social classes? Are the observed ethnic and racial differences merely a matter of the relative frequency in which certain genes are distributed in these groups? Or, inasmuch as the predominant majority of Indians, Mexicans, Negroes, and Puerto Ricans have lived since birth in conditions of poverty in families of little education and little hope, might one justifiably expect what they show on the tests and in the schools from the conditions of their rearing?

These questions are usually taken to involve the old issue of nature and nurture and to pose the question of their relative potency. Let me begin by recognizing that heredity is clearly primary. For such a matter as whether a given fertilized ovum becomes a human being or a member of another species, heredity is all important. One gets only elephants from breeding elephants, nothing else. Moreover, each individual begins life with a given complement of genes which he received in equal shares from each parent. The DNA in his genes carries information which exerts a continuing influence on his development

throughout his life. Since a life begins with conception, heredity is always primary. Yet, having such a primary status and exerting a continuing influence throughout life need imply neither genetically fixed traits nor a predetermined course and rate for later development.

Ways of Obtaining Answers

One approach to answering the questions raised here comes by way of definition. Cyril Burt (1969) has recently contended that intelligence should be defined as "innate, general, cognitive ability." Such a contention has been the dominant view in England ever since Sir Francis Galton (1869), Charles Darwin's younger cousin, conceived genetically fixed traits as implicit in the theory of natural selection, launched the study of individual differences with the publication of *Hereditary Genius*, and invented the correlation statistic to show the persistence of individual differences across generations. Karl Pearson, Galton's successor, improved on the correlation method, extended the scope of such investigation, and used the evidence to support his own such definition (Pearson, 1902, 1904).

Defining intelligence as innate ability has the defect of leaving it unmeasurable. Or, if scores on tests of intelligence are taken as measures of innate ability, this approach confuses an observable and measurable phenotype with the genotype which can only be inferred—a distinction which Johannsen (1903, 1909), one of the fathers of scientific genetics, made at the turn of the century. Ignorance or neglect of this distinction has all too often left psychologists discussing measures of intelligence as if they were genotypic. Moreover, the traditional distinction between tests of intelligence or aptitude from achievement tests tends to maintain such confusion. Here it has been the merit of Lloyd Humphreys (1962) to recognize that these two kinds of tests differ only in degree, with intelligence tests including a wider variety of items, tapping a wider variety of experience, and being further removed in time from relevant learning situations than achievement tests.

An approach by definition can start quite as justifiably with the contention that intelligence is primarily a product of learning rather than its cause. Thus, George Ferguson (1954, 1956) has explained abilities as derived from factor analysis as the results of transfer of training in overlearned skills. Moreover, Gagné (1968) has viewed

mental ability as a product of cumulative learning in which various skills form a transfer hierarchy ranging from stimulus-response connections through chains, motor and verbal, multiple discriminations, concepts, and rules, simple and complex. In the recognition of a hierarchical structure, Gagné's view resembles that of Piaget (1937, 1947), and my own (Hunt, 1961), yet Gagné's view gives less importance than either Piaget's or mine to the effect of the individual's existing organization of abilities on the consequences of his encounters with specified circumstances. Clearly, a definitional approach, by itself, can lead only to fruitless debate. The scores of all kinds of tests of abilities represent past developmental achievements in which the influence of the genes has interacted continuously with the influence of the environment in an on-going series of encounters.

The second approach is empirical. It is based upon the correlations between test scores from relatives, and the operations are quite independent of these belief-based definitions. Such correlations have been found regularly to be a positive function of the degree of genetic relatedness. Thus the correlations between monozygotic or identical twins are higher than those between dizygotic twins or siblings, or those between parents and children, which, in turn, are higher than those between cousins, and those between first cousins are somewhat higher than those between second cousins or unrelated children reared apart (see Erlenmeyer-Kimling & Jarvik, 1963; Fuller & Thompson, 1960, Ch. 7; Outhit, 1933). Moreover, the correlation between the scores for siblings reared apart ($r = +.34$ in Freeman, Holzinger, & Mitchell, 1928; median $r = +.47$ from 33 studies in Jensen, 1969, p. 49) is higher than that reported for first cousins ($r = +.06$ in Gray & Moshinsky, 1933; median $r = +.26$ from 3 studies in Jensen, 1969). Such findings and others (see Jensen, 1969) clearly attest a genetic influence on phenotypic measures of intelligence.

It has been more difficult to say how great this influence is. The traditional effort to separate and specify the relative strengths of the influences of heredity and environment has led to a variety of research and statistical designs beyond the scope of this essay (see Cattell, 1953, 1960; Falconer, 1960; Fuller & Thompson, 1960). The answer to the question has come in the form of estimates of heritability. Heritability is defined as that proportion of the variance within a specific population in the phenotypic measure of a characteristic that is determined by the genotypic variation within that population (Rieger, Michaelis, & Green, 1968, p. 213). In selection experiments, a simple approxima-

tion of heritability may be obtained by dividing the gain in the offspring by the selection differential. This gain is the difference of the mean of the trait measures concerned for the offspring from the mean for the population. The selection differential is the difference of the mean for those selected to be the parents from the mean of the population. The closer this ratio is to unity, the higher the heritability.

Since experimental selective manipulation of human matings has been out of the question for many reasons, estimates of heritability for human intelligence have been based upon statistical manipulations of the correlations between relatives. The findings have varied (see reviews by Erlenmeyer-Kimling & Jarvik, 1963; and by Fuller & Thompson, 1960). Galton (1883) began this effort with his studies of twins and concluded that "nature prevails enormously over nurture when the differences of nurture do not exceed what is commonly to be found among persons of the same rank of society and in the same county [p. 241]" although he lacked a numerical estimate. Such an estimate based upon the correlation between identical twins reared apart is conceptually the most obvious, if their environments are uncorrelated, because they have only their genes in common. Newman, Freeman, and Holzinger (1937) reported a correlation of +.77 between the Stanford-Binet IQs of their 19 pairs. More recently, according to Jensen (1969), Shields (1962) has also reported a correlation of +.77 between composite scores from 44 pairs on Raven's Progressive Matrices, and most recently, Cyril Burt (1966) has reported a correlation of +.86 between the Stanford-Binet IQs of 53 pairs of identical English twins. Such findings support both Galton's statement and the oft-quoted numerical estimate that 80 percent of phenotypic variance in the IQ derives from heredity.

Other bases for estimating heritability yield estimates differing little from the correlation between measures of intelligence for identical twins. The median of the correlations between such measures in five samples of unrelated children reared together is +.24. Since all these have in common is their environments, this serves as an estimate of variance in phenotypic intelligence due to environment. Subtracted from 100 percent, it leaves 76 percent due to heredity. Another basis comes from an attempt to assess the extent of family resemblance with cultural differences held constant by means of what was contended to be a culture-free test. Again, approximately 80 percent of the variance in phenotypic measures of intelligence among families was attributed to heredity (Cattell & Willson, 1938). When Jensen (1967) applied his

generalized formula for estimating heritability from any two correlation sets of relatives where the degree of kinship is higher for one than the other—to all those reported, he got a composite heritability value of +.77 (+.81 when corrected for unreliability). This he regards as the best overall estimate of the heritability of human intelligence (Jensen, 1969, p. 43). (Just how any single estimate can be expected for such a population statistic which varies with the variability of both genetic and environmental factors, he does not explain.)

Before World War II, textbooks of psychology often stated that heredity accounts for approximately 80 percent of the variance in the IQ. This is the assertion and the line of reasoning that I characterized, largely for purposes of exposition, as "the belief in fixed intelligence (Hunt, 1961)." Along with this belief went another that behavior unfolds directly with genetically controlled maturation which I termed "the belief in predetermined development" and which I traced back to G. Stanley Hall and his elaboration of what he saw as the implications of the recapitulation doctrine. This latter belief set the conceptual stage for the claim of a constant IQ and the notion that individual differences start with conception and remain essentially constant throughout development and life. It is basically this same argument which Arthur Jensen (1969) has revisited to explain what has come to be characterized as "the failure of Project Head Start." Implied also is that proposition that if 80 percent of the variance in phenotypic measures of intelligence derives from heredity, then most of those differences observed between the means of the IQs for children of professional and unskilled parents or the children of parents in the racial groupings must be biologically inevitable.

The Dissonant Evidence of Developmental Plasticity

This claim that 80 percent of the variance in phenotypic measures of intelligence and developmental advancement is attributable to heredity makes other evidence of developmental plasticity highly puzzling. Bits of such evidence began turning up even before World War II. They included the increases in the IQs of preschool children associated with nursery schooling in the Iowa studies (Skeels, Updegraph, et al., 1938; Wellman, 1940), lack of longitudinal predictive value for scores from tests given infants during their first two years even though these scores exhibited satisfactory reliability (Bayley, 1940), and the finding

by Skeels and Dye (1939) of dramatic increases in the IQs of orphanage-reared infants who were moved to an institution for the mentally retarded where "the older and brighter girls on the ward became very much attached to the children and would play with them during most of their waking hours" (p. 5; see Hunt, 1961). Instead of calling the beliefs in "fixed intelligence" and "predetermined development" into question, such was the firmness of faith in them that such evidences merely evoked a flood of methodological criticisms tending to discredit them.

More of such evidence has come since World War II. Spitz (1945, 1946a, 1946b) and Dennis (1960) reported apathy and dramatic retardation, even in locomotion, associated with orphanage rearing. Students of neonatal behavior turned up unsuspected evidences of ability for both classical and operant conditioning during the first few days following birth (for reviews see Lipsitt, 1963, 1966, 1967). Investigations stemming from the neuropsychological theorizing of Donald Hebb (1949) brought evidence that rearing chimpanzees in the dark hampers the development of their visually controlled behaviors (Riesen, 1947, 1958) and that increasing the complexity of early perceptual experiences of rats increases substantially their later ability to learn the Hebb-Williams (1946) mazes (Forgays & Forgays, 1952; Forgus, 1954; Hymovitch, 1952). According to this same principle, both rats (Hebb, 1947) and dogs (Thompson & Heron, 1954) which were pet-reared became significantly superior as adults to their cage-reared littermates in learning these mazes. For many skeptics, such evidences from the animal laboratory obviated selectivity and regression toward the mean, two of the criticisms commonly leveled against the evidence from the orphanage and nursery-school studies, as the basis for such early effects of experience on development. Thus, they lent also a cubit of credibility to the orphanage evidence.

Other evidence has suggested that the use of a sensorimotor organization solidifies its structure and hastens its development. In one of these studies, providing infants, beginning at five weeks of age, with a stabile pattern over their cribs to look at reduced the age at which the blink response appeared, to target drops of 11.5 inches, from 10.4 weeks of age for 10 control infants without such an opportunity to use their eyes, to an average of seven weeks for 10 infants provided with stabile patterns (Greenberg, Uzgiris, & Hunt, 1968). In another such investigation, providing infants with visual targets of complexity properly matched to their level of development reduced the age at

which they achieved mature reaching for a seen target from a median of 145 days, for infants in an original normative study, to a median of 89 days for the infants in a second enrichment study where the complexity of the visual stabile was properly matched to the development of the children (White, 1967). If one casts these findings into the terms of Stern's (1912) IQ ratio in order to put them in familiar perspective, lowering the age for the appearance of the blink-response from 10.4 weeks to 7 weeks is a gain of approximately 48 points, and lowering the achievement of that visuomotor coordination in mature reaching from a median of 145 days to a median of 89 days is a gain in the order of 63 points. Such a transformation applies only to past development and should imply no permanence unless the circumstances of these infants are so arranged as to give them special opportunities to accommodate their advanced visuomotor skills to new situations calling for further development.

Evidence of plasticity comes also for abilities more closely akin to tested intelligence and scholastic proficiency in what Piaget (1936, 1937) has termed object construction and imitation. We have developed a set of six sequentially ordinal scales (Uzgiris & Hunt, 1973). One is for object permanence and involves what is probably the most basic epistomological construction, an the other to be discussed here is for the development of vocal imitation.

Such ordinal scales enable one to define a level within a line of development existing between the top landmark passed and the first one failed. In cross-sectional studies, one can then compare the ages of children at given levels of development who are living under differing environmental conditions. In Athens are two orphanages for illegitimate babies – a municipal orphanage where the infant-caretaker ratio approximates 10/1, and Metera, which attempts to be a model baby center, where this ratio is of the order of 3/1. Whether an infant comes to one or the other of these orphanages appears to be a matter of chance. We have compared the ages of all the children from these two orphanages who were at the several levels on these scales of object permanence and vocal imitation, we also included a sample of children from working-class families. The results for object permanence appear in Figure 1. For illustrative purposes consider the means and standard deviations for that level of object permanence where a child follows an object through one hidden displacement but not through a succession of them. These appear in the left three of the middle cluster of vertical bars. For the Municipal Orphanage, the tallest bar on the

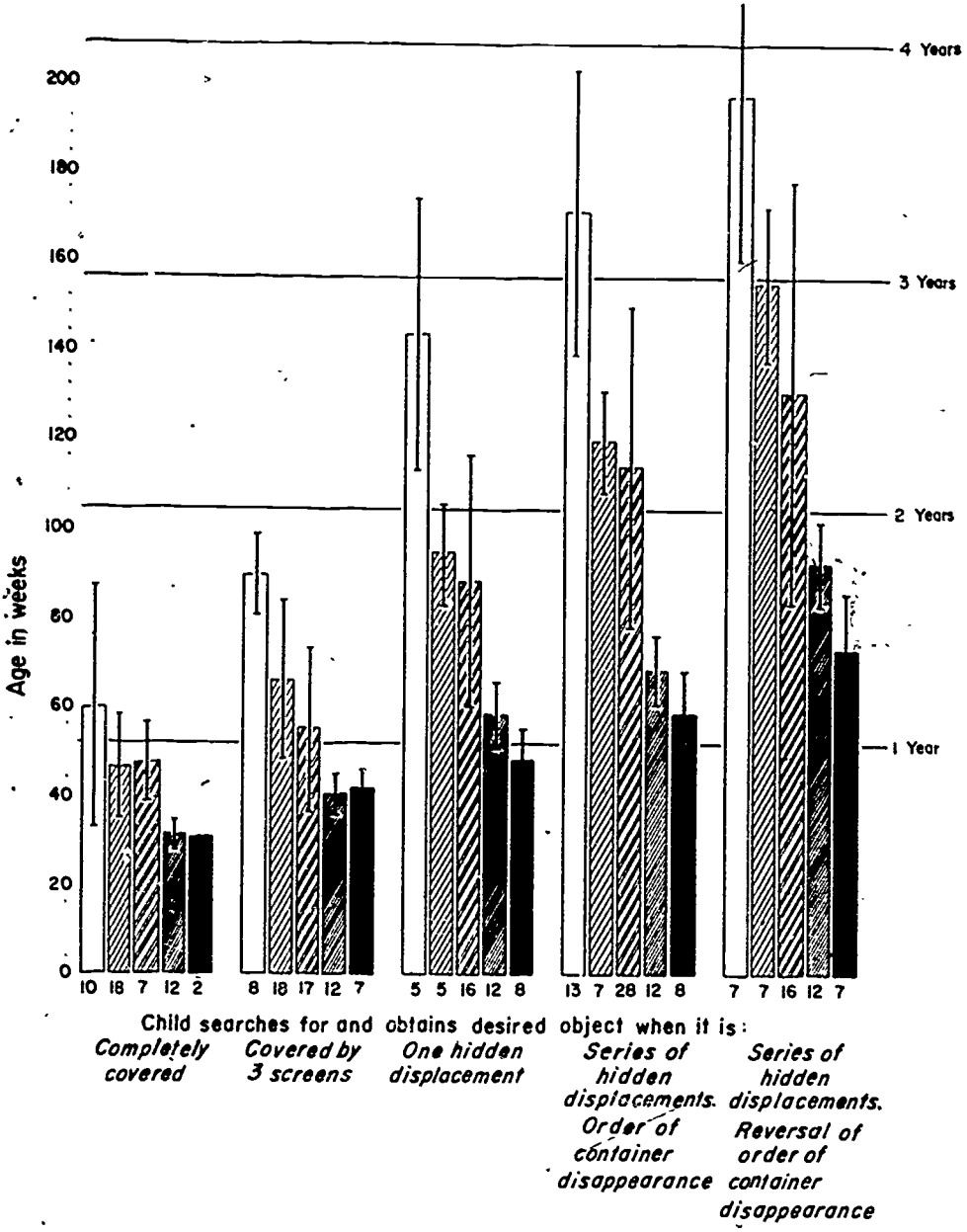


Fig. 1. Object construction under differing conditions of rearing.

left, the mean age for the five children at this level is 143.76 weeks with a standard deviation of 29.19 weeks; for Metera the mean, also for five children, is 94.39 weeks with a standard deviation of 11.13 weeks, and for the home-reared children, third bar, the mean age for the 16 children at this level is 87.9 weeks with a standard deviation of 28.06. The children of the Municipal Orphanage average significantly older than those from either Metera or the working-class homes who differ not significantly.

Consider also that level of object permanence at which a child follows an object through a series of hidden displacements and reverses the order in which the container of the desired object disappeared but fails to copy a series of five sticks decreasing in length from bottom to top. This is the top level on our scale of object permanence. For the Municipal Orphanage, the mean age for seven children at this level is 206.58 weeks with a standard deviation of 34.29; for Metera, the mean age of four children is 154.58 weeks with a standard deviation of 17.32 weeks; and for the 16 home-reared children at this level the mean age is 130.77 with a standard deviation of 47.11 weeks. Again, the children of the Municipal Orphanage are significantly older than either of the other groups who do not differ significantly. It is also worth noting the standard deviations. The smallest is that for Metera where the conditions of rearing are most nearly alike; next for the Municipal Orphanage where the child-caretaker ratio of 10/1 almost insures the choice of pets while others are neglected; but the standard deviation is largest by about 13 weeks for the home-reared children (Paraskevopoulos & Hunt, 1971). Presumably variations within families combine with differences in heredity to exaggerate the variance in the development of object permanence.

More directly relevant to the issue of the degree to which the class differences commonly observed are biologically inevitable is the evidence from two longitudinal studies made for quite other purposes. In one, a still unpublished study by Ina Uzgiris, the scales were administered every other week to 12 infants from middle-class homes. The other, also a still unpublished study by Schickedanz and myself, was made to evaluate the effects of a mothers' training program (Badger, 1971, 1972) on the development of the children of poverty being reared by parents participating in the Parent and Child Center of Mt. Carmel, Illinois. Here again, the scales were administered consecutively every other week to eight infants of these parents from the poverty sector who were participating in the program of this Center.

Let us consider the means and standard deviations of the ages at which these two samples of children achieved the same two levels of object permanence. In Figure 1, the results for these two samples appear in the two bars on the right of each cluster. For following a desired object through one hidden displacement, note again the middle cluster. The mean age for the children of middle-class homes was 58.46 weeks (S. D. = 7.43 weeks) and that of the infants of parents of poverty who were participating in the mother's training program was 48.06 weeks (S. D. = 9.22 weeks). All eight had achieved this level by one year of age, and the youngest by 44 weeks of age. For the top level of object permanence, following a desired object through a series of three hidden displacements in a reverse order, the mean age for those of middle-class was 91.36 weeks (S. D. = 9.43) and for the children of poverty in the Parent and Child Center was 72.74 weeks (S. D. = 13.99). Apparently this mothers' training program has hastened the development of object permanence. It is often assumed that the child-rearing of middle-class families approximates the optimum, but here is an instance in which the child-rearing of the parent caretakers from the poverty sector has been improved by a program of mother training to surpass that of the middle-class—at least for object permanence during the first two years of infancy. If support for the Parent and Child Center program is continued long enough, we shall be able to follow the development of these children and also their performance into school to see how their performance there compares with that of their older siblings who lived through their early years before their mothers were touched by Mrs. Badger's program of mother-training. We desperately need the kind of evidence one can get only from such prolonged longitudinal studies of children developing under differing environmental circumstances.

Only class differences have been concerned in these comparisons for object permanence. All of the children in both samples were white. But evidence of plasticity in phenotypic measures of intelligence comes also for black children from a demonstration underway in Milwaukee which is directed by Heber and Garber (1972), of the University of Wisconsin. According to a progress report (Heber, 1970) to the agency supporting their work, Heber and Garber started their investigation with a survey of tested intelligence in that census tract of Milwaukee with the lowest socioeconomic status. From this survey came the interesting finding that approximately 80 percent of those children with IQs below 80 came from families where the mothers

had IQs below 80. Since nearly all of the people in this survey were black, this finding fit well the hypothesis of biologically inevitable differences, but these investigators did not stop at this point. Rather, they selected a sample of 40 high-risk mothers with new infants. These mothers were identified objectively by IQs of 75 or below. For 20 of these families, a home visitor saw and played with each infant until the infant was approximately six months old, then Heber and Garber arranged to have the infant brought five days a week to a day-care center. There, each was cared for by a woman who had been selected for articulate speech and who had been trained to provide appropriate educational experiences for infants. Each infant was also put on a program of repeated testing, and the mother was given counseling in homemaking and in the buying and preparation of food. For the other 20 families, the program was limited to routine counseling visits with the mothers, and what I believe was a duplication of the program of repeated testing for their infants. At age 45 months, the IQs of the children of the control families average approximately 90. This is a whole standard deviation above that of their mothers, an unusual degree of increase which probably derives in part from the repeated testing as well as from the expected regression effect. At this same age of 45 months, the IQs of those children provided with educational day-care are reported to average an almost unbelievable 128. Unless there is something very wrong with this demonstration that I cannot now see, it provides spectacular evidence of plasticity in a standard phenotypic measure of intelligence within black children from families of the highest risk where mothers have IQs near the low end of the scale. Heber has been properly cautious about attributing permanence to such a gain. Moreover, I suspect considerable loss is inevitable if these children are simply returned to their families and to their neighborhood schools. Tests of intelligence measure only past achievements. They say very little of the future unless the circumstances of future development are specified.

MATURATION AND EXPERIENCE

In our traditional conception of development, maturation and learning have represented completely separate processes. Maturation has been considered to be controlled by the genes. Learning has been conceived to be controlled by environmental encounters. In this third quarter of the 20th century, however, clear evidence has come that informational

interaction through the eyes influences maturation within the central nervous system. For the most part, these investigations have been inspired by the neuropsychological theorizing of Donald Hebb (1949) and the neurobiochemical theorizing of Helgar Hydén (1959). Apparently inspired by Hebb's theorizing, Austin Riesen reared chimpanzees in the dark. The dark-rearing resulted not only in behavioral deficiencies but also in a diminution of the number of nerve cells and glial fibers developing within their retinal ganglia by adulthood (Chow, Riesen, & Newell, 1957). Corroboratively, Brattgård (1952), inspired by Hydén's biochemical theory of memory, reported that rearing rabbits in the dark results in a paucity of RNA production in their retinal ganglia as adults. Since then, a California group (Bennett, Diamond, Krech, & Rosenzweig, 1964; Krech, Rosenzweig, & Bennett, 1966) has reported that thickness of the cerebral cortex and the level of total acetylcholinesterase activity of the cortex, as well as rate of adult maze-learning, are the function of the complexity of the environment during early life. Quite recently, studies of the effects of dark-rearing during early life have been extended through the visual system. Wiesel and Hubel (1963), for instance, have demonstrated that dark-rearing produced a paucity of both cells and glial fibers in the lateral geniculate body of the thalamus, and a Spanish investigator, Valverde, and his collaborators, have obtained evidence that dark-rearing also decreases both dendritic branching and the number of spines which develop on dendritic processes of the large apical cells of the striate area in the occipital lobes in mice (Valverde, 1967, 1968; Valverde & Estebán, 1968). Evidence that dark-rearing diminishes higher-order dendritic branching in cats as well as mice has been reported by Coleman and Riesen (1968). Evidence suggesting that such effects on dendritic branching and spine density may be a matter of the complexity of the information encountered and the variety of adaptations called forth has come from studies by Holloway (1966) and Schapiro and Vukovich (1970). Most recently, Fred Volkmar and William Greenough (1972) have compared the dendritic branching of stellate and pyramidal neurones, Golgi stained, in several layers of the occipital cortex for litter-mate rats reared in a complex environment (where a group of 12 pups were housed in a large wire-mesh cage provided with a variety of toys that were changed daily), in environments consisting of pairs of animals in standard laboratory cages or single animals in such cages. Those reared in the complex environment exhibited considerably greater branching of dendrites of the third

order on. They showed seven times as many fifth-order branches on the pyramidal cells of the Layer V than did their litter-mates reared in isolation and about 3.5 times more than their litter-mates reared in social pairs. Since light was available for all, it would appear to be complexity of experience rather than mere absence or presence of light which is responsible for these very substantial differences. Not only do such findings show a great deal of environmental plasticity in the neuroanatomical maturation; they also suggest that the variety of informational inputs from circumstances of greater complexity call forth accommodative adaptions which show in neuroanatomy as well as in behavior. If one considers 80 percent of the variance in phenotypic measures of intelligence and related matters to be relevant to such evidences of plasticity in behavior and neuroanatomical maturation, then all this is highly puzzling, especially so since a partitioning of the variance in this study showed more than half of it related to environmental conditions (Volkmar & Greenough, personal communication).

The Norm of Reaction

Let us consider how much variations in environmental conditions appear from existing evidence to be able to alter phenotypic measures of intelligence.

Ever since Woltereck introduced the concept in 1909 (see Dunn, 1965), geneticists have concerned themselves with the "norm of reaction" or the "range of reaction" as well as with Mendelian statistics and the mechanisms of genetic transmission. The "norm of reaction" is defined as the range of phenotypic reactions which a specified genotype is able to produce in response to environmental influences (Rieger, Michaelis, & Green, 1968, p. 372). Such a concept, like that of heritability, can never be fully specified from empirical data because a new investigator with imagination can always arrange a new progression of environmental encounters which may alter further the range of phenotypic reactions. In investigative practice, however, one obtains relevant evidence in terms of the difference between the means of phenotypic measures of a given trait for samples of individuals from a given population of genotypes who are reared in different environments. For the complex trait of intelligence and scholastic ability, one can get evidence from comparing the mean values of phenotypic mea-

sures for individuals from a given population reared under differing environmental circumstances or differing educational programs.

Such data are still very few. In investigations with human subjects, moreover, they are seldom based strictly on a single population of genotypes. Nevertheless, highly suggestive evidence exists, and one set comes from the study of the ages at which children achieve the various levels of object permanence—perhaps the most basic of epistemological achievements. Here the Athenian children in the study by Paraskevopoulos and Hunt (1971) comprise about as close an approximation of a single population of genotypes as one can expect to get. They represent the lower half of the socioeconomic structure. The American children in the still unpublished Worcester study by Uzgiris represent the middle-class, but those from the Parent and Child Center at Mt. Carmel by Schickedanz and Hunt represent families of the lowest socioeconomic status who were recruited from those on Welfare and on Aid to Families with Dependent Children. For following a desired object through a series of hidden displacements in reverse order, the extreme mean ages are 206.58 weeks, for the children of the Municipal Orphanage where the child-caretaker ratio is 10/1, and 72.74 weeks for the children of poor families reared with the aid of the Badger Mother's Training Program in the Parent and Child Center at Mt. Carmel. For the Worcester children from middle-class families, the mean is 91.36 weeks with a standard deviation of 9.43 weeks. Even though I have been developing ordinal scales in part to escape Stern's IQs ratio and the normative approach to a meaning for test performances and scores (Hunt & Kirk, 1971), it may be worthwhile here for purposes of communication to cast these figures into this familiar ratio by assuming that 91.36 weeks for the children of middle class approximates the norm of 100. Thus, this empirical evidence indicates that the range of reaction must extend at least from a low of about 50 to a high of about 125—a range of reaction of 75 points of IQ.[2]

One may well object that such a range could be found for only a simple function during infancy when the longitudinal validity of measures of intellectual development is low. But this empirical difference

[2]The lower limit of this range deserves a word of explanation. Dividing 91.36 weeks by 206.58 weeks yields less than .5, but since the mean of 206.58 weeks derives from a cross-sectional study, it must exaggerate the delay more than would a longitudinal approach with examinations every other week. For this reason, I have rounded the lower limit to the approximation of an IQ of 50.

of 75 points is essentially the same as that found by Wayne Dennis (1966) when he had the Goodenough Draw-A-Man Test given to samples of typical children, aged between six and nine years, who were living in typical family environments in some 50 cultures over the world. The variation in the means of such IQs ranged from a low of 52 to a high of 124, a range of 72 IQ points. The variation in this phenotypic measure appears to be associated with the degree of contact with and participation in representative graphic art. Probably this Draw-A-Man Test calls for a considerably less complex set of abilities, as these are assessed by factor analysis, than either of the more standard scales: the Stanford-Binet test or the Wechsler-Bellevue Children's Scale. Yet, for American children, the IQs from the Draw-A-Man Test correspond about as well as do IQs from either of these other two measures with each other. It must also be admitted that children in a cross-cultural study cannot come from a single population of genotypes, yet typical individuals from the Syrian nomadic tribe have shown the capacity to adapt themselves to technological cultures when reared in them.

These empirically determined ranges of 72 and 75 points in mean IQs fall only about one standard deviation short of the full range of individual IQs (between 55 and 145) which includes all but a fraction of a percent of individuals above the pathological group which bulges at the low end of the distribution for the IQ. Clearly, there is dissonance between any argument based on the statistics of heritability and this argument from ranges of reaction.

Relevance: Heritability versus Norm of Reaction

Perhaps this dissonance can be clarified by the respective meanings of heritability and the norm of reaction. Heritability is, by definition, that proportion of the total phenotypic variance for a particular characteristic in a specified population (Rieger, Michaelis, & Green, 1968, p. 212). Heritability is not an attribute of a trait, but rather of a trait in a specified population developing and living within a given set of environmental conditions. What a given estimate of heritability gives is the amount of gain or loss to be expected in the course of selective breeding. Thus, given heritability for a particular trait in a given population at 80 percent, if a sample of parents is selected to have a mid-parent measure of the trait averaging one standard devia-

tion above the mean for the population, then the mean of the measures for that trait in the offspring would be expected to average .8 of a standard deviation above (or below) the mean of the population. This expectation would hold, however, if and only if the environmental circumstances remained constant through the lives of the two generations. Thus, an index of heritability tells us about how much of the selection advantage or disadvantage is lost between parents and their offspring.

On the other hand, an index of heritability can tell us nothing about how much change in the measures of a phenotypic trait will result from being reared in new environments.[3] It can tell us nothing about how much the IQs of the children from a given population of genotypes will be changed through being reared in newly designed environments and educational programs. Thus, a composite heritability index of 80 percent for the IQ may say how much of the variance in IQ is hereditary for the kinds of children studied who have developed under the existing environmental conditions of American and European culture, yet it tells us nothing of how much the IQ might be changed by newly designed systems of child-rearing and education. It is not relevant to why Project Head Start succeeded or failed. Knowledge of how much the IQ can be altered by new regimes of child-rearing and early education can be obtained only from evidences of the range of reaction or from studies of the differences between the means of IQ for children reared in differing environmental conditions and with differing programs of early education.

In a dynamic and developing society, the conditions of child-rearing and education are always changing, hopefully improving. Measures of phenotypic intelligence would be expected to go up with these supposed improvements unless the increases in intelligence are hidden in comparative scores based on new norms. They do. In a number of repeat studies, increases in average IQ have occurred rather than the loss predicted by Cattell (1937) from differential fertility (see Hunt, 1961). Cattell (1950) himself published one of these based on test surveys of the children in an English city in 1936 and 1939. Instead of the predicted drop of one point, he found a gain in mean IQ of 2.28 points.

[3]The dissonance between the evidences for a composite index of the order of 75 percent or 80 percent for heritability of the IQ and the evidences of plasticity in development has long puzzled me. I am greatly indebted to writings of Jerry Hirsch (1970, 1972) and to discussions with him for this clarifying interpretation of this dissonance.

Other studies have reported substantially larger gains in mean IQ. One by Smith (1942), based on surveys of the children aged 10 to 15 in the schools of Honolulu in 1924 and in 1938, reported a gain in mean IQ of 20 points. Another by Wheeler (1942), based on tests, made 10 years apart, of samples of students aged 10 to 12 years from given families in the schools of the Tennessee Valley before and after the changes introduced by the Tennessee Valley Authority, reported a 10-point gain in median IQ. Another by Finch (1946), based on tests given to all of the children in a sample of high schools in Minnesota during the 1920s and again in the 1940s, reported gains in means IQs ranging from 10 to 15 points. In yet another such study, Tuddenham (1948) compared a representative sample of draftees from World War II with a sample from World War I on comparable forms of the Army Alpha Group Test. The median for World War II fell at the 82nd percentile of the distribution of World War I. Thus, half of the draftees of World War II belong among the upper fifth for World War I (see Hunt, 1961, p. 337ff). So long as one considers a high index of heritability relevant to educability, such gains would seem incredible, but indices of heritability are not relevant.

Implications for Class and Race Differences

The implications of these considerations for class and race differences become readily apparent when one considers the inequalities of environmental opportunity across the class structure of American and European societies. Attempts to assess the genotypic potential behind phenotypic measures of intelligence have always assumed essentially equal environmental opportunity for growth, adaptation, and learning with micro-inequalities randomly distributed. The past two decades, however, have yielded abundant evidence of large deficiencies in the development-fostering quality of the environments provided for children of lower-class families of poverty whatever their ethnic origin and race. Inasmuch as a major share of Indian, Mexican, Puerto Rican, and black families fall in the poverty sector, their children share the poverty-based deficiencies of poor white families *plus* whatever additional disadvantages are associated with dark skins and differences in language.

These deficiences I have reviewed elsewhere in some detail (Hunt, 1969, pp. 202–214). They include basic nutritional deficiencies in a sub-

stantial share of mothers at the time of conception and during gestation. They include a lack of opportunities to acquire cognitive and linguistic skills illustrated by such facts as the following: that where approximately 90 percent of nursery school children can respond by picking up appropriate blocks for the colors named by an examiner and approximately 80 percent can name all of the six colors used, only approximately 20 percent of four-year-olds beginning a Head Start program can respond in these fashions (Kirk & Hunt, in preparation). These deficiencies also include a lack of opportunities to develop the motivational systems required for confidence and persistent striving and also the opportunities to acquire those values and standards of conduct demanded by the mainstream of a complex organized society. These are not small variations in environmental opportunity, and certainly they are not randomly distributed across the class structure.

Given the evidences of plasticity indicative of a range of reaction for measures of phenotypic intelligence of the order of 75 points, and given these class and race differences associated with poverty in the development-fostering quality of the environments, and especially the early environments provided for children, relatively small portions of the commonly reported deficits in the means of IQ and measures of scholastic performance for the children of unskilled parents—white, black, Indian, Mexican, and Puerto Rican—can be considered to be biologically inevitable. To be sure, the evidence and the argument summarized does not rule out a contribution from heredity to these differences. Yet if the Mother's Training Program of Earladeen Badger can bring the average IQ of even a small sample of families of poverty approximately 25 points above that for middle class families (as assessed by the scale of object permanence), and if the Heber-Garber program can bring the average Stanford-Binet IQ for children of black mothers with IQs of 75 or below up to 128 at age 45 months, it becomes hard to believe that more than a very minor share of the differences among class and ethnic groups are biologically inevitable.

Recently, this case against the biological inevitability of class and race differences has received empirical support from another direction. In a study reported at the Washington meetings of the American Psychological Association, George W. Mayeske (1971) reported a special analysis of the data in the report on *Equality of Educational Opportunity* by Coleman and others (1966) designed to ascertain the degree to which that 25 percent of the variance in scholastic achievement associated with racial and ethnic group membership could be

explained in terms of socioeconomic and educational circumstances. From the relevant partial regression equations, he took into account the socioeconomic status of each family, the presence or absence of key members of each family, the assessments of the aspirations for schooling by students and parents, parental beliefs about how students might benefit from education, their region and neighborhood of residence, and the achievement and motivations of the students attending the school. When this was done, the variance among the students in their academic achievement scores associated with ethnic-group membership dropped to 1.2 percent. Similar findings have been reported by Jane Mercer (1971) for Chicano and black children in the schools of Riverside, California. The more the families of these children resembled those of the modal configuration for the middle-class, white community of Riverside in terms of five characteristics, the more nearly did the mean IQs of the children approximate 100. In these studies, both the partial regression equations used by Mayeske and Mercer's modal characteristics of middle-class families contain potentially genetic variance, but the force of such a consideration is reduced by the evidence of IQs well above average for children of poor white families and of black mothers with low IQs when those children are provided with experiences which foster their psychological development. One would guess from such combinations of evidence that all but a very minor share of children of poverty or of unfavored ethnic and racial groups could be reared in a fashion which would permit them to perform quite adequately in our technological culture if the economy provided the opportunity. Moreover, many of those now typically fated for relative incompetence might well, with more fortunate rearing, achieve excellence along one of the diverse avenues of achievement in our society.

Recapitulation and Challenge

Let me recapitulate and then present what I see as the challenge. Significant deficiencies exist in the means of lower-class and certain racial groups for many measures of ability, motivation, and performance. Most of the evidence, however, concerns measures of intelligence and scholastic performance. Composite attempts to estimate the proportion of phenotypic variance in IQ which is genetic approximate 80 percent. These heritability indices have been interpreted to mean

that most of the observed deficiencies in the mean IQs for classes and races are biologically inevitable. These interpretations are in puzzling dissonance with evidences of plasticity which suggest that the range of reaction in the IQ must be of the order of 75 points. Such evidence of plasticity becomes less puzzling when one recognizes that estimates of heritability indicate the loss of deviation from the mean of the population to be expected from parents to offspring in experiments on selective breeding, that indices of heritability are relevant only to the status quo within a given population so long as environment remains constant, and that these indices say nothing about how much the mean of a phenotypic measure of intelligence or scholastic achievement will change with development in new environments. Such information demands knowledge of the norm of reaction which comes only from the difference between the means of measures of achievement and intelligence for groups of children from a given population of genotypes who have been reared under differing environmental circumstances or differing educational programs. Even though such evidence is sparse, that from two sources indicates that this range is of the order of 75 points and that special child rearing can boost the mean achievement for white children of poverty and for black children from mothers with IQs of 75 or below well above the population average.

If measures of heritability say nothing about educability, then the measures we have are irrelevant to the outcome of Head Start. But several factors help to explain why Head Start is said to have failed. I have elaborated these elsewhere (Hunt, 1969, Ch. 5). First, the goals were unrealistic, and, in terms of a broader view of social change, Head Start appears to have had considerable success, but in ways that differ from those unrealistic goals. Second, our basic understanding of psychological development and how to foster it was inadequate to the task. This explains in considerable part the unrealism of attempting to overcome the deficit deriving from four years of experience in a summer or a year of compensatory education without altering the milieu. Third, the nature of the nursery-schooling available for deployment in the crash program of Head Start was poorly adapted for the compensatory effect called for by the goals.

In one sense, the evidence outlined here may be viewed as optimistic, perhaps too optimistic. It is one thing to say that most of the class and race differences now evident are not biologically inevitable, and it is quite another to say that reducing the deficits associated with poverty is easy. We lack basic knowledge of early intellectual and

motivational development. Only recently have we begun to take seriously the hierarchical conception of such development and begun to describe the natural landmarks of achievement. An initial approximation of such landmarks exists now only for sensorimotor development (Uzgiris & Hunt, 1973) and for linguistic development during the preschool years (Brown, 1973). We know exceedingly little of what kinds of experience foster these successive landmarks and how each one is built upon earlier achievements. Even though we have a few instances where curricular suggestions have worked better than typical middle-class rearing, no suggestion we can make now is more than a hypothesis to be tested. Finding ways to get parents of poverty to utilize innovations in early education is another problem. From the experience of the Parent and Child Centers, it becomes clear that we have not even begun to evaluate programs in terms of the determinants of their success in eliciting parent cooperation and participation and to examine the factors responsible for success or failure. For many groups of parents any implied inferiority of their practices is hard to take, and resentment hampers cooperation with the program. Discovering ways of harnessing class, neighborhood, and ethnic pride to get parental cooperation in the improvement of early education along with discovering the kinds of experience which foster early intellectual and motivational development are the basic challenges for the behavioral and social sciences and professions concerned in early childhood education.

Questions and Answers

Q: To what age does plasticity continue? Can twelve-year-olds catch up? Or, does the work you reported mean that trying to help high school students is useless?

A. Although the evidence appears to indicate that plasticity decreases considerably with age and with the diminishing rate of anatomical maturation, at least some degree of plasticity continues even to senility or perhaps death. Insofar as a person my age can have his theoretical views modified by evidence or even learn a set of nonsense syllables, a degree of plasticity remains intact. At age twelve, the investment of effort required for a major modification of abilities or motivational system appears to be substantially greater than it would

be at ages four or five, and greater yet than it would be at birth. This diminution of plasticity appears to be due not only to diminution in the rate of maturation, but also to the increasing levels of ability already acquired, and to the amount of information in the storage. In terms of the hierarchical conception of the achievements included in what we measure with tests of intelligence, individuals who fall behind have typically failed to develop abilities, conceptions, and motivational systems that enable them to process new information and cope with new situations.

Now, is attempting to help high school children useless? Of course it is not useless. On the other hand, it is likely to require more investment of time and effort to achieve that minimum of competence and motivation required to participate in the mainstream of our complex technological culture than would compensatory education at ages four and five. Moreover, such compensatory education can be expected to call for more investment in this sense than would be required for infants and young children from birth on. Modifying the child-rearing of parents, however, is not without its own difficulties.

Q: To what extent does nutrition affect brain-cell growth and, therefore, intelligence? Can nutritional differences account for observed class-related differences?

A: Reports by Cravioto, by MacDonald, and by Pasamanick have reported that the incidence of dietary deficiencies and chronic health problems is about four or five times as high among the families of the poor as among families with average or higher incomes. The literature on the association between neural maturation and nutritional deficiencies is growing. Unfortunately, I do not know this literature well enough to be able to synopsize it with confidence. I know better the literature on the role of informational interaction in neural maturation, and I have synopsized some of this. From what I do know about the literature, it would appear that nutritional factors are among those associated with class-related differences in intelligence as now measured.

Q: What evidence exists that performance on object permanence is related to more widely used measures of intelligence?

A: Nancy Bayley has included performances on object permanence in her infant scales. Piagetians and cognitive theorists very commonly

believe that the ontogenesis of object construction is not easily modified by experience. Our data show that the range of reaction for the ages at which infants and toddlers living under differing environments achieve the highest level on our scale of 14 steps in object construction extends from sixteen and a fraction months to approximately 45 months. Since object construction is among the most "intellectual" of the performances used in any infant scales, where motor behavior is usually used, it should be one of the best indicators of intellectual achievement. I know, however, of no correlational studies of the relationship of object construction to the more widely used tests of intelligence. On the other hand, later indices of intellectual development from Piaget have shown high correlations with results from standard tests of intelligence.

Q: Do you know of any evidence bearing upon the improvement in the IQ of high SES children through enrichment programs? Might not class differences be maintained if these are also "plastic?"

A: It is generally assumed that the child-rearing of middle class families approximates the optimum. If this should be the case, children in middle class families should be developing about as rapidly and as fully as their genotypes permit. I doubt seriously if this is the case. On the other hand, their developmental rates probably more nearly achieve their genotypic limits than do the developmental rates of the children of poverty. The discussion of this matter is amplified in my written paper.

Q: Will these miraculously raised IQs in infants and young children endure at older age levels?

A: This is a serious question. From the standpoint of theory, however, one would expect these raised levels to endure only if the circumstances fostering the elevation of the IQ at the early ages persisted through the later ages and stages of development. The traditionally accepted constancy of the IQ appears to be a function of at least three quite different factors: first, the genotypic variations in learning potential traditionally assumed, second, the progressively decreasing part-whole ratios involved in test-retest consistencies with age; and third, the consistency in the development-fostering quality of the circumstances encountered within most families and most neighborhoods. Leon Yarrow has found measures of characteristics in familial

environments based upon two three-hour time-samplings accounting for between 20 and 25 percent of the variance in measures of performance on the Bayley scales at six months. This finding suggests that the consistency in the development-fostering quality of environments within families and neighborhoods is far greater than we have imagined, and far more important for the observed constancy in the IQ than we have ever believed. As a converse corollary, the evidences of plasticity indicated by the available evidences of the range of reaction tend to confirm this importance.

Q: In view of your discussion of mother-child-IQ relationships and your reference to the work of Heber at the University of Wisconsin, what do you think of the recent court decision in Iowa where children were removed from one home because of parental incompetence, the ascertainment of which was based on a low IQ?

A: The case in question is unknown to me. I would contend that there are more relevant indicators of competence for child-rearing than an IQ. In Puerto Rico, for instance, Dr. Albizu-Miranda has reported instances of children no more than seven or eight years old having higher mental ages than their parents, yet those parents were managing their lives adequately. On the other hand, if the Heber finding that a very high proportion of children with intellectual and motivational defects come from a relatively small number of families where the parents are intellectually or motivationally incompetent turns out to be reproducible, then his other finding that educational day-care for the children of such parents can bring their competence up to or beyond average suggests that we may have to supplement what parents can be taught and what they can do for their young in these relatively few cases.

REFERENCES

Anastasi, A. *Differential psychology. Individual and group differences in behavior.* (3rd ed.) New York: Macmillan, 1958.

Badger, E. A mother's training program—the road to a purposeful existence. *Children*, 1971, **18** (5), 168–173.

Badger, E. A mother's training program—a sequel article. *Children Today*, 1972, **1** (3), 7–12.

Battle, E. S., & Rotter, J. B. Children's feelings of personal control as related to social class and ethnic group. *Journal of Personality*, 1963, **31**, 482–490.

Bayley, N. Mental growth in young children. *Yearbook, National Society for Studies in Education*, 1940, **39**, (2), 11–47.

Bennett, E. L., Diamond, M. C., Krech, D., & Rosenzweig, M. R. Chemical and anatomical plasticity of the brain. *Science*, 1964, **146**, (No. 3644), 610–619.

Brattgård, S. O. The importance of adequate stimulation for the chemical composition of retinal ganglion cells during early postnatal development. *Acta Radiologica*, 1952, Stockholm, Suppl. 96, 1–80.

Brown, R. W. The development of the first language in the human species. *American Psychologist*, 1973, **28** (in press).

Burt, C. The genetic determination of differences in intelligence. A study of monozygotic twins reared together and apart. *British Journal of Psychology*, 1966, **57**, 137–153.

Burt, C. What is intelligence? *British Journal of Educational Psychology*, 1969, **39**, 198–201.

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

Cattell, R. B. The fate of national intelligence. Test of a thirteen-year prediction. *Eugenics Review*, 1950, **42**, 136–148.

Cattell, R. B. Research designs in psychological genetics with special reference to the multiple variance method. *American Journal of Human Genetics*, 1953, **5**, 76–91.

Cattell, R. B. The multiple abstract variance analysis equations and solutions: For nature-nurture research on continuous variables. *Psychological Review*, 1960, **67**, 353–372.

Cattell, R. B., & Willson, J. L. Contributions concerning mental inheritance. I. Intelligence. *British Journal of Educational Psychology*, 1938, **8**, 129–149.

Chow, K. L., Riesen, A. H., & Newell, F. W. Degeneration of retinal ganglion cells in infant chimpanzees reared in darkness. *Journal of Comparative Neurology*, 1957, **107**, 27–42.

Clark, R. E. Psychoses, income, and occupational prestige. *American Journal of Sociology*, 1949, **54**, 433–440.

Coleman, J. S., et al. *Equality of educational opportunity*. Washington, D. C.: U. S. Government Printing Office, 1966.

Coleman, P. D., & Riesen, A. H. Environmental effects on cortical dendritic fields. I. Rearing in the dark. *Journal of Anatomy*, 1968, London, **102**, 363–374.

Dennis, W. Causes of retardation among institutional children: Iran. *Journal of Genetic Psychology*, 1960, **96**, 47–59.

Dennis, W. Goodenough scores, art experience, and modernization. *Journal of Social Psychology*, 1966, **68**, 211–228.

Dunn, L. C. *A short history of genetics*. New York: McGraw-Hill, 1965.

Erlenmeyer-Kimling, L., & Jarvik, L. F. Genetics and intelligence: A review. *Science*, 1963, **142**, 1477–1479.

Falconer, D. S. *An introduction to genetics*. New York: Ronald Press, 1960.

Faris, R. E. L., & Dunham, W. *Mental disorders in urban areas*. Chicago: University of Chicago Press, 1939.

Ferguson, G. A. On learning and human ability. *Canadian Journal of Psychology*, 1954, **8**, 95–112.

Ferguson, G. A. On transfer and the abilities of man. *Canadian Journal of Psychology*, 1956, **10**, 121–131.

Finch, F. H. Enrollment increases and changes in the mental level of the high school population. *Applied Psychology Monograph*, 1946, **10**, 75.

Forgays, D. G., & Forgays, J. W. The nature of the effect of free environmental experience in the rat. *Journal of Comparative and Physiological Psychology*, 1952, **45**, 322–328.

Forgus, R. H. The effect of early perceptual learning on the behavioral organization of adult rats. *Journal of Comparative and Physiological Psychology*, 1954, **46**, 331–336.

Freeman, F. N., Holzinger, K. J., & Mitchell, B. C. The influence of environment on the intelligence, school achievement, and conduct of foster children. *Yearbook of the National Society for Social Studies of Education*, 1928, **27** (Part 1), Chap. 9.

Fuller, J. L., & Thompson, W. R. *Behavior genetics* New York: Wiley, 1960.

Gagné, R. M. Contributions of learning to human development. *Psychological Review*, 1968, **75**, 177–191.

Galton, F. *Hereditary genius. An inquiry into its laws and consequences*. London: Macmillan, 1869.

Galton, F. *Inquiries into human faculty and its development*. London: Macmillan, 1883.

Gough, H. G. The relationship of socio-economic status to personality inventory and achievement test scores. *Journal of Educational Psychology*, 1946, **37**, 527–540.

Gough, H. G. A new dimension of status. II. Relation of the St scale to other variables. *American Sociological Review*, 1948, **13**, 534–537.

Gray, J. L., & Moshinsky, P. Studies on genetic psychology. The intellectual resemblance of collateral relatives. *Proceedings of the Royal Society: Edinborough*, 1933, **53**, 188–207.

Greenberg, D. J., Uzgiris, I. C., & Hunt, J. McV. Hastening the development of the blink-response with looking. *Journal of Genetic Psychology*, 1968, **113**, 167–176.

Hebb, D. O. The effects of early experience on problem-solving at maturity. *American Psychologist*, 1947, **2**, 306–307.

Hebb, D. O. *The organization of behavior*. New York: Wiley, 1949.

Hebb, D. O., & Williams, K. A method of rating animal intelligence. *Journal of Genetic Psychology*, 1946, **34**, 59–65.

Heber, R. F. *Rehabilitation of families at risk for mental retardation: A progress report*. Madison, Wisconsin: Research and Training Center in Mental Retardation, University of Wisconsin, 1970.

Heber, R., & Garber, H. Familial risk factors in mental retardation. *Human Differences and Social Issues*, 1972 (in press).

Hertzig, M. E., Birch, H. G., Thomas, A., & Mendez, O. A. Class and ethnic differences in the responsiveness of preschool children to cognitive demands. *Monographs of the Society for Research in Child Development*, 1968, **33**, (1).

Hirsch, J. Behavior-genetic analysis and its biosocial consequences. *Seminars in Psychiatry*, 1970, **2** (1), 89–105.

Hirsch, J. Genetics and competence: Do heritability indices predict educability? In J. McV. Hunt (Ed.), *Human intelligence*. New Brunswick, N. J.: Transaction Books, 1972. Pp. 7–29.

Holloway, R. L. Dendritic branching—Some preliminary results of training and complexity in rat visual cortex. *Brain Research*, 1966, **1**, 393–396.

Humphreys, L. G. The nature and organization of human abilities. In M. Katz (Ed.), *The 19th Yearbook of the National Council on Measurement in Education*. Ames, Iowa, NCME, 1962. Pp. 39–45.

Hunt, J. McV. *Intelligence and experience*. New York: Ronald Press, 1961.

Hunt, J. McV. *The challenge of incompetence and poverty. Papers on the role of early education*. Urbana: University of Illinois Press, 1969.

Hunt, J. McV., & Kirk, G. E. Social aspects of intelligence: Evidence and issues. In R. Cancro (Ed.), *Intelligence*. New York: Grune & Stratton, 1971, Pp. 262–306 (Chap. 15).

Hydén, H. Biochemical changes in glial cells and nerve cells at varying activity. In F. Brücke (Ed.), *Proceedings of the 4th International Congress on Biochemistry*. Vol. 3. *Biochemistry of the central nervous system*. London: Pergamon, 1959. Pp. 64–89.

Hymovitch, B. The effects of experimental variations in early experience on problem solving in the rat. *Journal of Comparative and Physiological Psychology*, 1952, **45**, 313–321.

Jensen, A. R. Estimation of the limits of heritability of traits by comparison of monozygotic and dizygotic twins. *Proceedings of the National Academy of Science*, 1967, **58**, 149–157.

Jensen, A. R. How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 1969, **39**, 1–123.

Johannsen, W. *Ueber Erblichkeit in Populationen und in reinen Linien*. Jena: Fischer, 1903.

Johannsen, W. *Elemente der exakten Erblichkeitslehre*. Jena: Fischer, 1909.

Krech, D., Rosenzweig, M. R., & Bennett, E. L. 1966. Environmental impoverishment, social isolation, and changes in brain chemistry and anatomy. *Physiology and Behavior*, **1**, 99–104 (Pergamon Press, Great Britain).

Landis, C., & Page, J. D. *Modern society and mental disease*. New York: Farrar & Rinehart, Inc., 1938.

Lipsitt, L. P. Learning in the first year. In L. P. Lipsitt & C. C. Spiker (Eds.). *Advances in child development and behavior*. Vol. 1. New York: Academic Press, 1963. Pp. 147–195.

Lipsitt, L. P. Learning processes of human newborns. *Merrill-Palmer Quarterly*, 1966, **12**, 45–71.

Lipsitt, L. P. Learning in the human infant. In H. W. Stevenson, R. Hess, & H. L. Rheingold (Eds.), *Early behavior. Comparative and developmental approaches*. New York: Wiley, 1967. Pp. 225–247.

Mayeske, G. W. 1971. On the explanation of racial-ethnic group differences in achievement test scores. Paper presented at the Meetings of the American Psychological Association, Washington, D. C., September 1971. Washington, D. C.: Office of Education.

McNemar, Q. *The revision of the Stanford-Binet Scale*. Boston: Houghton-Mifflin, 1942.

Meichenbaum, D. H., & Goodman, J. Reflection-impulsivity and verbal control of motor behavior. *Child Development*, 1969, **40**, 785–798.

Mercer, J. R. Pluralistic diagnosis in the evaluation of black and Chicano children: A procedure for taking sociocultural variables into account in clinical assessment. Paper presented at the Meetings of the American Psychological Association in Washington, D. C., September 1971. Riverside, California: Department of Sociology, University of California.

Mumbauer, C. C., & Miller, J. O. Socioeconomic background and cognitive functioning in preschool children. *Child Development*, 1970, **41**, 471–480.

Newman, H. H., Freeman, F. N., & Holzinger, K. J. *Twins: A study of heredity and environment*. Chicago: University of Chicago Press, 1937.

Outhit, M. C. A study of the resemblance of parents and children in general intelligence. *Archives of Psychology*, 1933, **149**, 1–60.

Paraskevopoulos, J., & Hunt, J. McV. Object construction and imitation under differing conditions of rearing. *Journal of Genetic Psychology*, 1971, **119**, 301–321.

Pearson, K. The law of ancestral heredity. *Biometrika*, 1902, **2**, 211–228.

Pearson, K. On the inheritance of mental and moral characters in man. *Biometrika*, 1904, **3**, 23–237.

Piaget, J. 1936. *The origins of intelligence in children*. (Margaret Cook, Transl.) New York: International Universities Press. 1952.

Piaget, J. 1937. *The construction of reality in the child*. (Margaret Cook, Transl.) New York: Basic Books, 1954.

Piaget, J. 1947. *The psychology of intelligence* (M. Piercy & D. E. Berlyne, Transls.) Paterson, N. J.: Littlefield, Adams & Co., 1960.

Rieger, R., Michaelis, A., & Green, M. M. *A glossary of genetics and cytogenetics. Classical and molecular*. (3rd ed., English) New York: Springer-Verlag, 1968.

Riesen, A. H. The development of visual perception in man and chimpanzee. *Science*, 1947, **106**, 107–108.

Riesen, A. H. Plasticity of behavior: Psychological aspects. In H. F. Harlow & C. N. Woolsey (Eds.), *Biological and biochemical bases of behavior*. Madison: University of Wisconsin Press, 1958. Pp. 425–450.

Schapiro, S., & Vukovich, K. R. Early experience effects upon cortical dendrites: A proposed model for development. *Science*, 1970, **167**, 292–294.

Shaw, C. R. *Delinquency areas*. Chicago: University of Chicago Press, 1929.

Shields, J. *Monozygotic twins brought up apart and brought up together*. London: Oxford University Press, 1962.

Short, J. F. Jr., & Strodtbeck, F. L. *Group process and gang delinquency*. Chicago: University of Chicago Press, 1965.

Shuey, A. M. *The testing of Negro intelligence*. (2nd ed.) New York: Social Science Press, 1966.

Sims, V. M. Relations between the social class identification and personality adjustment of a group of high school and college students. *Journal of Social Psychology*, 1954, **40**, 323–327.

Skeels, H. M., & Dye, H. B. A study of the effects of differential stimulation of mentally retarded children. *Proceedings of the American Association on Mental Deficiency*, 1939, **44**, 114–136.

Skeels, H. M., Updegraff, R., Wellman, B. L., & Williams, H. M. A study of environmental stimulation: An orphanage preschool project. *University of Iowa Study of Child Welfare*, 1938, **15**, No. 4.

Smith, S. Language and non-verbal test performance of racial groups in Honolulu before and after a fourteen year interval. *Journal of Genetic Psychology*, 1942, **26**, 51–93.

Spitz, R. A. Hospitalism. An inquiry into the genesis of psychiatric conditions in early childhood. *The Psychoanalytic Study of the Child*, 1945, **1**, 53–74.

Spitz, R. A. Hospitalism: A follow-up report. *The Psychoanalytic Study of the Child*, 1946a, **2**, 113–117.

Spitz, R. A. Anaclitic depression. *The Psychoanalytic Study of the Child*, 1946b, **2**, 313–342.

Stern, W. 1912. *The psychological methods of testing intelligence* (G. M. Whipple, Transl.). Baltimore: Warwick & York, 1914.

Thompson, W. R., & Heron, W. The effects of restricting early experience on the problem-solving capacity of dogs. *Canadian Journal of Psychology*, 1954, **8**, 17–31.

Tuddenham, R. D. Soldier intelligence in World Wars I and II. *American Psychologist*, 1948, **3**, 149–159.

Uzgiris, I. C., & Hunt, J. McV. Toward ordinal scales of psychological development in infancy. Unpublished manuscript, University of Illinois, Champaign, 1973.

Valverde, F. Apical dendritic spines of the visual cortex and light deprivation in the mouse. *Experimental Brain Research*. 1967, **3**, 337–352

Valverde, F. Structural changes in the area striata of the mouse after enucleation. *Experimental Brain Research*, 1968, **5**. 274–292.

Valverde, F., & Esteban, M. E. Peristriate cortex of mouse: Location and the effects of enucleation on the number of dendritic spines. *Brain Research*, 1968, **9**, 145–148.

Volkmar, F. R., & Greenough, W. T. Rearing complexity affects branching of dendrites in the visual cortex of the rat. *Science*, 1972, **176** (4042) 1445–1447.

Warner, W. L., Meeker, M., & Eells, K. *Social class in America. A manual of procedure for the measurement of social status*. Chicago: Science Research Associates, 1949.

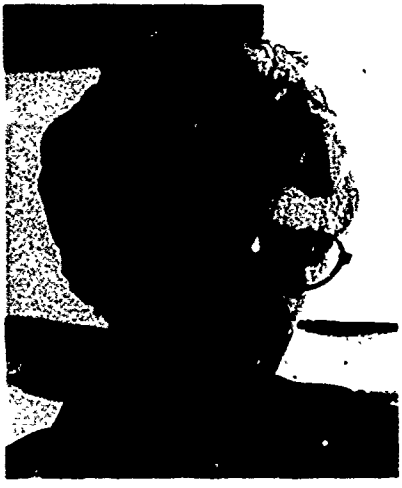
Wellman, B. L. Iowa studies on the effects of schooling. *Yearbook of National Sociological Studies of Education*. 1940, **39**, 377–399.

Wheeler, L. R. A comparative study of the intelligence of East Tennessee mountain children. *Journal of Educational Psychology*, 1942, **33**, 321–334.

White, B. L. An experimental approach to the effects of experience on early human development. In J. P. Hill (Ed.), *Minnesota Symposia on Child Development*. Minneapolis: University of Minnesota Press, 1967 Pp. 201–226.

Wiesel, T. N., & Hubel, D. H. Effects of visual deprivation on morphology and physiology of cells in the cat's lateral geniculate body. *Journal of Neurophysiology*, 1963, **26**, 978–993.

Zigler, E., & Butterfield, E. C. Motivational aspects of changes in IQ test performance of culturally deprived nursery school children *Child Development*, 1968, **39**, 1–14.



Eleanor E. Maccoby

Sex Differences In Intellectual Functioning[1]

ELEANOR E. MACCOBY
and
CAROL NAGY JACKLIN
Stanford University

This paper is in the nature of a progress report. We are currently in the process of preparing a new edition of the 1966 book, *The Development of Sex Differences*, and are reviewing all the relevant literature we can locate on the topic of this paper as part of the process of updating the annotated bibliography for that book. Our work is not yet completed, and at this time we can only say what some of the trends appear to be. I would like to begin by discussing whether data available during the past seven years would cause us to reconsider any of the generalizations that we and others had come to in 1966.

First, the performance of the two sexes on measures of total, or composite abilities, such as IQ tests: it is still a reliable generalization that there are no sex differences on these tests. The question itself is not a very interesting one, however. As we all know, boys are better at certain kinds of items and girls at others, and the sexes can be made to differ on one direction or the other, or to be the same, according to the choice of a particular mix of items for a test. Since time is limited today, I would prefer to move directly to the differences in components.

[1]Revised version of a paper by the same name in *The Development of Sex Differences*, Stanford University Press, 1966.

Differences in Verbal Ability

Recent research continues to find female superiority on a range of verbal tasks. However, it may be that we need to reconsider some of our views concerning the course of development of this difference with age. It has been thought that verbal differences begin very early—from the time of the utterance of the first word or even earlier, in babbling, but that the sexes become more alike on verbal skills as they grow older. If we go back to the source of this generalization, the 1954 McCarthy summary of studies of language development, we find that the differences reported tended to be small, and many, as McCarthy noted, were not significant even on large samples. It was just that when there was a difference it favored girls, and the many studies taken together added up to a significant trend. The same was true generally in the studies done between the McCarthy study and our own 1966 review (Maccoby, 1966), although the study with the largest sample, Templin (1957), found no sex differences over the age range three to six. In fact, however, there has been almost no work with children under two and one-half or three of a normative sort, involving large and unspecialized samples of children, since the 1930s and 1940s. Work in the field of language development has been very intensive during the past fifteen years, and our understanding of this development has been very greatly enhanced, but the work has tended to be focused upon very small and rather highly selected groups of children. The fact is that we do not know whether there has been a change in the relative standing of the two sexes at these early ages with respect to articulation, length of utterance, or early vocabulary. Small-scale recent studies, such as those of Reppucci (1970) with two-year-olds, and Roberts and Black (1972) with children of one and one-half to two, have found no sex differences. Beginning at the age of two and one-half, we do have some recent work with large samples. McCarthy and Kirk (1963) tested children ranging from two and one-half to nine to obtain norms for the Illinois Test of Psycholinguistic Abilities (ITPA). They found no consistently significant sex differences. The only consistent trend across age levels was that boys were better at "visual decoding"—that is, at *receptive* naming when the stimulus was visual. A set of seven other recent studies involving children of pre-school age has found no sex differences on a variety of verbal tasks. A major exception is in the work done by LTS with children from impoverished families. Here, girls are ahead on productive language, though a

difference is not found on "receptive", or "passive" language—that is, on the understanding of words spoken by others.

As we move into the next age range, the early school years, there are again few differences. Brimer (1969), who gave vocabulary tests to very large samples indeed, found, in fact, higher average scores for boys at each age from six through 11. Most studies, however, including the ITPA norming sample mentioned above, detect no consistent sex differences, and these include tasks involving productive "fluency" as well as tests of understanding. The primary exception is found in the work of the Stanford Research Institute, with very large samples of disadvantaged children in Follow-Through programs from kindergarten through the second grade. Here the girls clearly test higher in a variety of language skills including reading, vocabulary, and the understanding of relational terms.

It is at about age 10 or 11 that girls begin to come into their own in verbal performance. It is from this age through the high school and college years that we find them outscoring boys at a variety of verbal skills. Sex differences are not found in every study; the findings seem to depend in part upon whether tests of general knowledge are called verbal tests—boys tend to do at least as well as girls on such tests, and in the Project Talent sample, substantially better. But in tests of verbal power, girls above age 11 do better, and in some studies the difference is fairly large in absolute terms. We have expressed as a rough estimate that, during adolescence, girls score on the average about a quarter of a standard deviation higher than boys on verbal tasks. One longitudinal study (Droege, 1967) which followed a large group of high school students from the ninth to the 12th grade found that the superiority of girls on verbal tasks increased through this period. This study is especially interesting since its longitudinal design permitted a control for differential dropout. We think it is important to be clear that we are not talking only about spelling, punctuation, and talkativeness. Included as well are considerably higher-level skills, such as comprehension of complex written text, quick understanding of complex logical relations expressed in verbal terms, and verbal creativity of the sort that is measured by Guilford's tests of divergent thinking.

We suggest that there are distinct phases in the development of verbal skills in the two sexes through the growth cycle. One occurs very early—before the age of three. We emphasize that the studies documenting sex differences at this age are very old, and that we do not know that the same situation prevails today. If it does, the evi-

dence indicates that the differences lie exclusively in productive language, not receptive language, and the girls' advantage is short-lived. At about three the boys catch up, and in most population groups the two sexes perform very similarly until adolescence. The exceptions during this phase are found in populations of underprivileged children, where girls maintain an advantage to a later age. We suggest that boys' greater vulnerability to hazards of all sorts, including those prevailing prenatally, means that the worse the prenatal and postnatal nutrition and medical care prevailing in a population, the greater the sex difference in early performance will be and the higher the age to which the difference will persist, because of the presence of larger numbers of low-scoring boys who have suffered some sort of systemic damage in the populations most at risk. We will return shortly to the matter of variability and its possible causes; but now let us simply note that for large unselected populations the situation seems to be one of very little sex difference in verbal skills from about three to 11, with a new phase of differentiation occurring at adolescence.

Math Ability

The earliest measures of some aspect of quantitative ability begin at about age three with measures of number conservation, soon followed by enumeration. There appear to be no sex differences in performance in these tasks during the preschool years, nor in mastery of numerical operations and concepts during the early school years, except in disadvantaged populations. Here again the data from the large studies conducted with Head Start and Follow-Through children show the girls to be ahead. The majority of studies on more representative samples show no sex differences up to adolescence, but when differences are found in the age range nine to 13, they tend to favor boys. After this age, boys tend to move ahead, and the sex differences become somewhat more consistent from one study to another, though there is great variation in the degree of male advantage that is reported. For example, Project Talent finds that boys' math scores are two-thirds of a standard deviation better than girls' at the 12th grade, while Droege, also using thousands of cases, finds no significant sex difference in high school, a large Swedish study finds a difference of less than one-fifth standard deviation. It is not possible to estimate, at this point, how large quantitative differences are likely to be.

Spatial Ability

Spatial ability continues to be the area in which the strongest and most consistent sex differences are found. But the superiority of boys emerges relatively late. In the tests of psychomotor skills administered during the first two years of life, the sexes do not differ on items which might be thought to have a spatial component (form boards, for example). There are now a number of studies in which modified versions of the Embedded Figures Test have been given to children of preschool age. One set of studies finds no sex differences (Reppucci, 1970; Shipman, 1971; Sitkei & Meyers, 1969; Eckert, 1970; Lewis et al., 1968). Coates (1972) found preschool *girls* to be superior on embedded figures, and Corey (1970), working with large numbers of children in kindergarten and first grade, found girls to be superior on "geometric design" while boys did better on mazes. Sex differences remain minimal and inconsistent until approximately the age of 10 or 11, when the superiority of boys becomes consistent on a wide range of populations and tests.

Before we consider some possible explanations of the trends we have described, let us consider what is known concerning variability in the performance of the two sexes at successive ages.

Variability

The question of whether boys are more variable in their intellectual performance than girls has been with us for a long time and still is not solved. The question was raised initially in Terman's work, when he identified more boys as "gifted (Terman, 1925)." The excess of boys having IQs over 140 was found on a test where there were no sex differences in the means of large samples, hence it appeared that boys must be more variable, including more of both unusually high scorers and unusually low scorers. As Miles and Terman both noted in the 1954 edition of the Carmichael Manual (Miles, 1954), the method of selection of cases for the Terman study made interpretation of the sex ratios difficult. The initial screening for location of high-scoring children was done with children who were nominated by their teachers as being especially bright, plus some additional children who volunteered for the testing. We know that girls tend to underestimate their own intellectual abilities more than boys do, and so there is danger of sex

bias in testing self-selected groups. Both Miles (1954) and Terman and Tyler (1954) reviewed a number of studies to find out whether there was a concentration of either sex among the very high scorers on tests of mental abilities. They concluded that there appeared to be no consistent tendency toward a higher incidence of giftedness among boys, and that the sex ratios in the gifted range depended on the content of the test.

Considering the mean sex differences reported earlier on verbal, spatial, and mathematical tests, it should come as no surprise that there would be a higher incidence of very high-scoring boys on tests which emphasize content in which boys, as a group, do better. In some recent tests to identify children with extremely high math and science abilities at the junior high school level, Julian Stanley and his associates have located considerably more boys than girls with these talents. Presumably, if one looked for the exceptionally high scorers on verbal tests, one would find more girls. Of course such results do not necessarily mean that one sex is more variable than the other – the two distributions could have equal standard deviations but with the distribution of one sex simply displaced upward, yielding more cases above any arbitrary cutting point.

At the lower end of the ability scale, studies of the incidence of learning deficits consistently indicate that there are more boys than girls who suffer from such deficits. The greater vulnerability of the male child to anomalies of prenatal development, birth injury, and childhood disease is well known and needs no documentation here. For our present purposes let us simply say that this vulnerability probably does affect the incidence of very low scores on tests of mental abilities. In school systems, children with such scores tend to be siphoned off into classes for the educationally handicapped, and in most psychometric work done with school children, these classrooms are not included. Their inclusion would, of course, increase the variability of boys' scores by adding more scores at the low end of the distribution. When these cases are excluded, however, the distributions of the two sexes on a variety of tests are remarkably similar. We have plotted standard deviations for the two sexes, by age, on tests of verbal, spatial, and quantitative abilities. Up to adolescence, we find no tendency for either sex to have larger standard deviations. After age 11, boys' standard deviations tend to be about five to six percent higher than the girls'. Our tabulations are not completed, but our conclusion at present is that there is very little sex difference in variability

prior to adulthood. The patterns of abilities which occur among the gifted members of the two sexes may be expected to differ somewhat, but there appears to be no basis for expecting a difference in the number of boys and girls who have very high over-all intellectual power. That men have historically achieved greater eminence in science, literature, and the arts we do not doubt. What we do doubt is that this difference is rooted in a greater incidence of very high intellectual potential during the adolescent years.

Possible Origins of Intellectual Sex Differences

It is time now to consider possible causes of the sex differences that have been outlined briefly above. We will begin by examining possible genetic components, and will consider how genetic factors might be carried and expressed during development. Then we will turn to experiential—environmental—factors.

We all know some of the standard research on heritability of the IQ. Composite intelligence does seem to have a substantial heritability factor. When we turn to the question of inheritance of specific abilities, particularly those which may be sex-linked, the genetics become more complex, and we find little research that is directly relevant. We have been able to locate four studies of parent-child resemblances in spatial ability (Stafford, 1961; Corah, 1965; Hartlage, 1970; Bock & Kolakowski, 1973). All show significant cross-sex correlations. That is, boys' scores on tests of spatial abilities are correlated with their mothers' scores but not their fathers'. Girls' scores are correlated with their fathers' but less with their mothers'. Stafford's hypothesis is that at least one important genetic determiner for spatial ability is sex-linked, being carried on the X chromosome, and being recessive. Girls, with their two X chromosomes, would have a relatively low chance of receiving two recessives, which would have to exist for the trait to be manifest. Whenever boys got a recessive X, it would be manifest, since there would be no dominant X to suppress it. Furthermore, boys would always get their X from their mothers, not their fathers. The magnitudes of the correlations that have been found in the three studies do fit the Stafford hypothesis reasonably well. It should be noted that these parent-child studies virtually rule out the possibility that spatial ability is acquired through same-sex modeling. The modeling hypothesis is implausible at the outset, because of the

fact that there is not much about an individual's spatial thinking that is open to the observation of others. What precisely is a child to copy if he is to learn spatial thinking by imitation? The overt responses involved in, say, the solution of a Gottschaldt figures problem tell very little about the thinking that led to the solution. Despite this problem, it has been suggested that since men on the average have higher space ability, boys must acquire it by selective imitation of their fathers rather than their mothers. The studies of parent-child resemblance show that children in a given family are more likely to resemble the cross-sex parent, and that if boys have higher space ability this occurs despite their similarity to their mothers, not because of any matching to their fathers' scores.

It appears likely, then, that there is at least some degree of genetic control over spatial ability. We have not yet located comparable figures for verbal abilities, so we do not know whether girls' verbal superiority has a similar basis. But where does our genetic knowledge about spatial ability leave us? Knowing that there is a genetic component in this ability does not help much in understanding the process and events that are involved in the development of spatial thinking. Does the space gene control the rate of development of certain parts of the brain? Does the space gene operate through the mediation of sex hormones?

An intriguing notion has recently been advanced by Kimura (1967), by Levy-Agresti and Sperry (1968), and others, that space ability is related to the degree of lateralization between the two hemispheres of the brain. As it bears upon sex differences, one line of reasoning is as follows: in adults, spatial ability tends to be localized in the right hemisphere of the brain, verbal functioning in the left. In early childhood, the two hemispheres of the brain are not greatly specialized in function, and lateralization tends to occur over a period of growth. Girls are on a faster developmental timetable than boys. Hence hemispheric dominance for some functions tends to be established earlier for girls. Evidence for this comes from work by Kimura using a dichotic listening technique. If different verbal messages are presented to the two ears, the adult listener tends to hear what came to the right ear, since this message goes mainly to the left side of the brain where speech is localized. In testing middle-class sample children aged five through eight, Kimura found this speech lateralization well established in children of both sexes through the whole age range tested. But later work with economically less-advantaged children

showed that five-year-old girls had established left-hemisphere specialization for speech while five-year-old boys had not (Kimura, 1967). And there is evidence that in boys with reading disabilities, the lag in lateralization is even greater (Taylor, 1962). Another instance of sex differences in rate of development of cerebral dominance is provided by Ghent (1961), who reports that differential touch sensitivity between the preferred hand and the non-preferred hand normally has developed in girls by age six and in boys it is delayed till age 11.

So far we seem to have a possible explanation of the boys' early lag in verbal functioning and their catch-up in middle childhood, with the lag continuing longer in children from disadvantaged populations. If we can assume lateralization to be an advantage for most intellectual functions, however, the rate-of-maturation hypothesis would help to explain male deficits early in life, but they would not explain female deficits. And the greatest problem with the hypothesis is that, as we have seen, there is really very little sex difference in any of the major component abilities during early and middle childhood. Sex differences emerge strongly after the age of 11, when lateralization is presumably as complete as it is going to be in both sexes. It is possible, of course (as Sperry has suggested) that once left-hemisphere dominance has been established it tends to inhibit the development of functions that would normally be specializing in the minor hemisphere. Thus, the boys' delay in establishment of hemispheric dominance might give them time for visual-spatial functions to develop strongly. However, they do ultimately develop left-hemisphere dominance for speech functions, so the question remains why does male spatial development show the greatest spurt during adolescence? Of course, there may be delayed effects. To find out whether early language development in any way shuts off the development of spatial ability, what is needed, we think, is examination *within* sex, of the relationship between the rate of early language development and later levels of spatial ability. If boys who talk late have better spatial ability in later childhood than boys who talk early, this would be good evidence for the inhibiting effects of early left-hemisphere dominance on right hemisphere functions.

In the absence of this kind of data, we are pretty much in the dark as to whether there is an early critical period for spatial ability when the absence of lateralization is especially important.

Levy-Agresti and Sperry present a different line of reasoning with respect to brain dominance. They argue that lateralization is less

strong in women, and in fact that women and left-handed men share a number of problems related to weak hemisphere dominance, particularly deficits in the visual-spatial sphere. The Levy-Agresti and Sperry hypothesis might explain the large and increasing superiority of males on spatial tasks, but it will not explain the large and increasing female superiority in verbal and conceptual thought. We find the Levy-Agresti and Sperry hypotheses and the Kimura hypothesis incompatible. That is, one cannot explain male spatial ability as due to stronger lateralization in males, and then attempt to explain female verbal superiority as due to stronger (or earlier) lateralization in females. Both things are currently being said.

Another problem that fosters skepticism about the brain-lateralization explanation of sex differences is the nature of the cluster of skills controlled by the two hemispheres. The right hemisphere controls spatial visualization, at which girls are worse, but it also controls fine perceptual-motor coordination, at which girls are better; the left hemisphere controls language fluency and reading, at which girls tend to be better; but it also controls other elements where girls have no advantage. Levy-Agresti (1968) has described the left hemisphere as follows: it is verbal, sequential, detailed, analytic, and computer-like. This does not sound like the traditionally feminine package of skills—or even like a psychological package we can associate with any group!

We are prepared to believe that differential brain lateralization may eventually turn out to be related to the different patterns of abilities the two sexes develop. But precisely what the relationships are seem very unclear at the moment, and we must simply consider it an open question.

Let us turn now to a different mechanism through which a genetically-controlled sex difference might work. We refer to the action of male and female hormones. We might begin with the well-known study by Ehrhardt and Money (1967), in which a group of ten girls were located whose mothers had been given an androgenic hormone during pregnancy. The ten girls had all been physically masculinized to some degree. Their social behavior and interests showed masculine tendencies as well. For our present purposes, the point of major interest is that the girls at age 13-14 had an average IQ of 125. Ehrhardt and Money note that this is above the national norm of 100. It should be noted, however, that these girls came from families with higher than average education. Six out of nine of the fathers had college education. One does not know precisely what level of IQ to expect from daughters

of untreated mothers in families at this education level, but it is safe to say that the average would be above 100. More interesting is the work by Dalton (1968), in which pregnant women were given dosages of progesterone, a *female* hormone. Follow-up work has been done with both the male and female offspring of these women. The children have significantly higher IQ scores than matched controls, and this is true of both sons and daughters born following progesterone-treated pregnancies. It appears that either male or female hormones can serve to promote whatever aspects of prenatal growth relate to ultimate intellectual strength. But the work so far does not help us to understand the development of different patterns of abilities in the two sexes.

There are similar problems with the work of Broverman and his colleagues (Broverman, 1964; Broverman et al., 1968; Klaiber et al., 1971). Broverman and others classify abilities in a different way than along spatial or verbal dimensions; they claim that males are better at tasks that call for the inhibition of previously-learned responses, and those that call for restructuring, while females are better at simple, overlearned tasks. In reviews of this work Parlee (1972) and Kagan and Kogen (1970) raised questions about the validity of this classification, and we will not repeat their criticisms here except to say that we believe they are cogent. For our present purposes, the point of greatest interest is the effects of experimentally administered sex hormones on performance in certain selected tasks. Unfortunately, the researchers used only male subjects and used only serial subtraction as a task for measuring the effects of the hormone administration. If serial subtraction can qualify as a simple overlearned task, then it would be included with the group of feminine skills. When testosterone was administered, performance on this task declined less, over an interval of time, than when no testosterone was administered, suggesting that the male hormone facilitated performance on this feminine task. This finding is consistent with the Broverman view that male and female hormones act in the same direction – that is, toward feminization of intellectual performance – but that female hormones are more powerful. Unfortunately the experimental design does not permit a test of this view, since female hormones were not administered, and no task requiring inhibition of a previously learned response was used. Thus we cannot compare the effects of the two kinds of hormones on what Broverman considers to be typically masculine and typically feminine tasks. It would be highly useful, in research of this kind, if both male and female subjects were used, since there might easily be an interac-

tion between the individual's hormonal history and the effects of an added hormone.

At this point, we must simply say that we consider the work on the relation between sex hormones and intellectual performance to be totally inconclusive.

Having encountered a good deal of frustration in our efforts to understand the possible biological factors underlying sex differences in intellectual performance, let us now briefly consider the social-emotional and experiential factors. Remembering that the major sex differences emerge after the age of 11, we are inclined to give a good deal of weight to the social-emotional influences that impinge primarily at adolescence. We believe it is true that many girls do develop what Horner has called the "will to fail" at this age. Patterns of male dominance and female acquiescence probably also become especially strong in the interactions between the two sexes at this age, and bright girls do begin to hesitate to compete with boys, or to take opposing positions in their presence. The problem is simply this: we can detect a number of factors which might have a generally inhibiting effect on the intellectual growth of one sex or the other. What we have to explain, however, is an increasing superiority of girls on verbal tasks, and an increasing superiority of boys in spatial tasks and possibly some quantitative tasks as well. What social factors improve intellectual performance of one kind while interfering with other kinds? We are hard put to guess what those factors might be. We have not been able to locate any solid research which relates any sort of social pressure or parental socialization practices in adolescence to specific patterns of abilities. Of course we may be dealing with delayed-action effects; it might be that there is something about the ways boys and girls are treated in early childhood that predispose them to certain patterns of intellectual development in adolescence. Witkin, for example, has argued that maternal behavior which fosters independence in young boys will be associated with a boy's developing a high level of "perceptual differentiation" a characteristic which has a high loading on spatial ability as it is ordinarily measured. This might help to explain the later emergence of high spatial ability in boys if we could document that they are granted greater independence during childhood. A recent survey we have done on differential parent treatment of the two sexes during the first six years of life (Maccoby, 1972) has indicated that the two sexes are about equally restricted during this period—girls are allowed as much independence as boys, on the aver-

age. We also have reason to doubt that there are clear differences in dependency between the two sexes during these early years, at least as this variable is usually measured. Furthermore, mothers talk to their young sons as much as they do to their young daughters. The notion that girls' superiority in verbal development is rooted in greater early childhood dependency, or greater amounts of verbal stimulation or reinforcement, simply cannot be documented from the evidence we have at present on parent-child interaction. What about the social milieu in which girls and boys live during middle childhood and late adolescence? Here there may be, and probably are, some differences in how much freedom of movement and freedom from surveillance is allowed the two sexes, but we lack evidence. There can be no doubt that boys spend more time on sports, and that girls are more interested in other kinds of social interaction. But let us think seriously about the kinds of tasks that girls do well on during the high school years: tests of verbal abilities at this age include verbal analogies, selection of precise opposites for relational terms, and logical problem solving. Is a girl going to be able to solve an item on the Miller Analogies Test, or make a correct deduction from a premise in a syllogism item, because she has chattered with her girl friends after school instead of playing ball? Is she going to be better at diagramming sentences simply because she has uttered more sentences during the day? We are profoundly skeptical about hypotheses of this kind. We suspect that the girls who spend most time chatting about nonschool-related matters are not going to be the girls who will get the highest scores on tests of abstract verbal ability, and that girls' greater interest in social affairs is not the explanation of their superiority on such tests. We encounter the same kind of problem when we attempt to explain boys' spatial superiority in terms of their greater interest in sports, or other activities that would mean more moving about through space. It seems unlikely to us on the face of it that boys will be helped in developing spatial ability by throwing a basketball through a hoop while girls are not helped by threading a needle; yet this is about the level of much current speculation on the subject. Empirically, not much is known concerning the characteristics of high-space boys, but Witkin's work suggests that they tend to be rather quiet and bookish. In a study of differential abilities done at Stanford some years ago (Ferguson & Maccoby, 1966), we found that high-space boys were significantly lower in aggression, and significantly more withdrawn, than low-space boys. Thus the kinds of interpersonal activities that boys do engage in

more frequently than girls, such as rough and tumble play, do not seem to foster the development of the ability in which they most excel: spatial ability. In fact, it would appear that they have good spatial ability in spite of their patterns of daily activities, not because of them.

We feel we should apologize for having given you a recital of what we do *not* know about the origins of intellectual sex differences. We would like to have been able to be more positive. But perhaps divesting ourselves of some misconceptions may not be a bad way to begin the complex task of understanding the factors that underlie sex differences in intellectual functioning.

Questions and Answers

Q: Is there any information on the possible effects of emotional and attitudinal differences between the sexes on subsequent development of sex differences in specific abilities?

A: There have been efforts to link some of the well-documented emotional sex differences, such as the greater aggressiveness of the male, to specific intellectual abilities. There is a psychoanalytic hypothesis, for example, that mathematics favors males because it is essentially aggressive, that is, in such mathematical analysis as that involved in algebra, for example, one must continually destroy existing statements and substitute new ones. To call such mathematical operations "aggressive" seems to us inherently implausible, but if one were to take the hypothesis seriously, one would have to account for male superiority on *non*-aggressive (and hence feminine") aspects of mathematics such as integral calculus.

There is little doubt that the sexes do differ from an early age in certain temperamental qualities, boys *are* more competitive, girls more conforming. Furthermore, the massive input of sex hormones at the beginning of adolescence does change the internal emotional climate for both sexes, a fact which might well have implications for their intellectual development during this period. There is very little research on what the implications are, however. In speculating about promising possibilities, we recommend taking a cue from the work of Schachter. He found that an injection of a given hormone would serve as a general arouser, but that the nature of the emotional state which

was experienced and expressed depended on aspects of the external situation. In a similar vein, we regard it as likely that both male and female hormones function as arousers. The two kinds of hormones may have sex-specific effects, of course, on the behavior of adolescent boys and girls toward members of the opposite sex. When it comes to their effects on intellectual interests and performance, however, we suggest that an influx of sex hormones might make the individual either more effective or self-defeating, and strengthen tendencies in either an anti-intellectual or a pro-intellectual direction, depending upon the individual's social milieu and his habitual modes of dealing with increased activation. In short we believe that the effects of sex hormones on intellectual performance, though potentially powerful, will prove to be susceptible to channeling through social influence.

Q: What are the implications of early sex-typing in interests and activity preferences for the later development of specific intellectual abilities?

A: There is good evidence for sex-typing in toy preferences during the preschool years, and for considerable differences in the extracurricular interests of boys and girls during middle childhood. We have commented in our paper that we do not see boys' interest in sports as directly linked to spatial or mathematical ability. Boys' interests in mechanical gadgets of all kinds, however, may be important precursors of these abilities. To our knowledge, the longitudinal evidence to support or refute this hypothesis does not exist. There is some recent evidence that girls at about the age of 10 begin to believe that studying math and science will have little relevance to their future work or life styles, while boys *do* see these studies as relevant. The emergence in early adolescence of views about the probable nature of their future lives may have a good deal to do with the rapid set differentiation in intellectual interests and skills that occurs at this age.

REFERENCES

Bock, D. R., & Kolakowski, D. Further evidence of sex-linked major-gene influence on human spatial visualizing ability. *The American Journal of Human Genetics*, 1973 (in press).

Brimer, M. A. Sex differences in listening comprehension. *Journal of Research and Development in Education*, 1969, 3 (1), 72–79.

Broverman, D. M. Generality and behavioral correlates of cognitive styles. *Journal of Consulting Psychology*, 1964, 28, 487–500.

Broverman, D. M., Klaiber, E. L., Kobayashi, Y., & Vogel, W. Roles of activation and inhibition in sex differences in cognitive abilities. *Psychological Review*, 1968, 75, 23–50.

Coates, S. W. *P.E.F.T. Preschool Embedded Figures Test*. Palo Alto: Consulting Psychologists Press, 1972.

Corah, N. L. Differentiation in children and their parents. *Journal of Personality*, 1965, 33, 301–308.

Corey, M. T. Use of Wechsler preschool and primary scale of intelligence in admissions testing and for educational placement. Paper presented at the American Psychological Association, Miami, Florida, September 1970.

Dalton, K. Ante-natal progesterone and intelligence. *British Journal of Psychiatry*, 1968, 114, 1377–1382.

Droege, R. C. Sex differences in aptitude maturation during high school. *Journal of Counseling Psychology*, 1967, 14, 407–411.

Eckert, H. M. Visual motor tasks at 3 and 4 years of age. *Perceptual and Motor Skills*, 1970, 31, 560.

Ehrhardt, A. A., & Money, J. Progestin-induced hermaphroditism: IQ and psychosexual identity in a study of ten girls. *The Journal of Sex Research*, 1967, 3 (1), 83–100.

Ferguson, L. R., & Maccoby, E. E. Interpersonal correlates of differential abilities. *Child Development*, 1966, 37, 549–571.

Ghent, L. Development changes in tactual thresholds on dominant and non-dominant sides. *Journal of Comparative and Physiological Psychology*, 1961, **54**, 670–673.

Hartlage, L. C. Sex-linked inheritance of spatial ability. *Perceptual and Motor Skills*, 1970, **31**, 610.

Kagan, J., & Kogan, N. Individuality and cognitive performance. In P. Mussen (Ed.), *Manual of child psychology*, Vol. 1. New York: Wiley, 1970. Pp. 1273–1366.

Kimura, D. Functional asymmetry of the brain in dichotic listening. *Cortex*, 1967, 3, 163–178.

Klaiber, E. L., Broverman, D. M., Vogel, W., Abraham, G. E., & Cone, F. L. Effect of infused testosterone on mental performances and serum LH. *The Journal of Clinical Endocrinology and Metabolism*, 1971, **32**, 341–349.

Levy-Agresti, J. Ipsilateral projection systems and minor hemisphere function in man after neocomissurotomy. *Anatomical Record*, 1968, **61**, 1151.

Levy-Agresti, J. & Sperry, R. W. Differential perceptual capacities in major and minor hemispheres. Paper presented at fall meeting, National Academy of Sciences, November 1968, California Institute of Technology, Pasadena, California. *Proceedings of the National Academy of Sciences*, 1968, **61**.

Lewis, M., Rausch, M., Goldberg, S., & Dodd, C. Error, response time and IQ: Sex differences in cognitive style of preschool children. *Perceptual and Motor Skills*, 1968, **26**, 563–568.

Maccoby, E. E. Differential socialization of boys and girls. Paper presented at the American Psychological Association meeting, Honolulu, Hawaii, September 1972.

Maccoby, E. E. Sex differences in intellectual functioning. In Eleanor E. Maccoby (Ed.), *The development of sex differences*. Stanford, Calif.: Stanford University Press, 1966.

McCarthy, D. Language development in children. In L. Carmichael (Ed.), *Manual of child psychology* (2nd ed.) New York: Wiley, 1954. Pp. 492–630.

McCarthy, J J. & Kirk, S. A. *The construction, standardization and statistical characteristics of the Illinois Test of Psycholinguistic Abilities*. University of of Illinois, Institute for Research on Exceptional Children, 1963.

Miles, C. C. Gifted children. In L. Carmichael (Ed.), *Manual of child psychology*. (2nd ed.) New York: Wiley, 1954. Pp. 984–1063.

Parlee, M. B. Comments on the "roles of activation and inhibition in sex differences in cognitive abilities" by D. M. Broverman, E. L. Klaiber, Y. Kobayashi, and W. Vogel. *Psychological Review*, 1972, **79**. 180–184.

Project Talent. *Counselors' technical manual for interpreting test scores*. University of Pittsburgh and American Institute for Research, 1961, P. 44.

Reppucci, N. D. Parental education, sex differences, and performance on cognitive tasks among two-year-old children *Developmental Psychology*, 1970, **4**, 248–253.

Roberts, G. C., & Black, K. N. The effect of naming and object permanence on toy preferences. *Child Development*, 1972, **43**, 858–868.

Sherman, J. A. *On the psychology of women: A survey of empirical studies*. Springfield, Ill : Charles C. Thomas, 1971.

Shipman, V. C. Structure and development of cognitive competencies and styles prior to school entry. Princeton, N J.: Educational Testing Service, 1971, PR-71-19.

Sitkei, E G. and Meyers, C. E. Comparative structure of intellect in middle- and lower-class four-year-olds of two ethnic groups *Developmental Psychology*, 1969, **1**, 592–604.

Stafford, R. E. Sex differences in spatial visualization as evidence of sex-linked inheritance. *Perceptual and Motor Skills*, 1961, **13**, 428.

Svensson, A. *Relative achievement—school intelligence, sex and home environment*. Stockholm: Almqvist & Wiksell, 1971.

Taylor, L. B. Perception of digits presented to right and left ears in children with reading difficulties. Paper read at meeting of Canadian Psychological Association, Hamilton, Canada, 1962

Templin, M. C. *Certain language skills in children—their development and interrelationships*. Institute of Child Welfare Monograph No. 26. Minneapolis: The University of Minnesota Press, 1957.

Terman, L. M. *Genetic studies of genius*. Vol. 1. Stanford, Calif.: Stanford University Press, 1925.

Terman, L. M., & Tyler, L. E. Psychological sex differences. In L. Carmichael (Ed.), *Manual of child psychology*. (2nd ed.) New York: Wiley, 1954. Pp. 1064–1114.

Witkin, H. A., Dyk, R. B., Faterson, H. F., Goodenough, D. R., & Karp, S. A. *Psychological differentiation*. New York: Wiley, 1962.

Implications of Group Differences for Test Interpretation

LLOYD G. HUMPHREYS
University of Illinois
Urbana

My topic concerns the sources of individual and group differences in test scores for cognitive traits and the implications of these differences for test interpretation. Actually the topic could be broadened to include any quantifiable information concerning human abilities or achievements, including interviews, letters of recommendation, and other ratings, with little if any modification of the discussion. The problem generally is one of drawing inferences about an individual from any source of information or combination of sources, whether the information be categorical or continuous. Nevertheless, I shall use "test" for convenience, but please remember the broader context and the broader implications of my discussion.

My discussion also centers on normatively scaled tests administered for purposes of revealing individual differences in performance. Failure to discuss such tests as mastery, criterion referenced, and diagnostic does not mean that these other types of tests are unimportant. To the extent that good tests of these latter types are available, they should certainly be used to supplement, but not to supplant, the information obtained from normative tests of abilities and achievements.

There has been, historically, considerable concern about the "fairness" of the inferences drawn from test scores for individuals from certain subgroups in our population. Twenty years ago interest centered on lower-class children. Today the focus is on ethnic or racial

minorities, both children and adults. Fairness, however, is a word laden with overtones of past and current wrongs, discrimination, and so on, and I shall for the time being substitute accuracy of inference for fairness of inference.

The accuracy of an inference is assessed along two classical dimensions: the amount of random error present and the amount of constant error present. Criterion measures, like predictor information, can be either continuous or categorical, and the method of assessment of the accuracy of the inference differs slightly in the two situations. For a continuous criterion, the standard error of estimate reflects random error; for a categorical criterion the analogue is the size of the difference between the conditional probabilities relating test score to the two or more categories. Constant error is assessed by the amount of over- or underestimation of the score on the continuous criterion, or of the over- or underestimation of the conditional probabilities for the categorical criterion, as a function of the demographic group to which the individual belongs.

When there is an independent, continuous criterion measure available, these two dimensions of accuracy of a test are translatable into the regression comparison described by Gulliksen and Wilks (1950) and popularized by Cleary (1968). The size of the standard error of estimate is inversely related to the slope of the regression line, and the assessment of constant error involves comparing the intercepts of parallel regression lines or of differences between nonparallel lines. Comparisons of conditional probabilities, including both differences among categories and among groups, are equivalent statistical operations for independent categorical criterion measures.

Without an independent criterion measure on individuals, which is the usual situation in the use of academic achievement tests, inference is made to a theoretical true score. Since the true score is theoretical, it can always be considered continuously and normally distributed. In this situation, random error is directly related to measurement error and constant error is determined by judging the degree of content validity of the test in two or more groups. In doing the latter it is right and proper to compare test content to the curriculum content to which the examinee group has been exposed. To compare the test to the content of the total life learning experiences of the examinee group subverts the purpose in giving an achievement test.

I am now ready to define the fairness of an inference concerning an individual drawn from test scores. Random error should be minimized

within reasonably generous limitations of testing time, and the amount and direction of the constant error should be known and allowed for.

The first aspect of fairness means that validity of the test, or combination of tests, should be as high as possible within broad practical limitations. There is rarely any good excuse for making inferences affecting the lives of individuals on the basis of very short tests, or shortened forms of standard tests. If the inference is worth making, in fairness to the individual, it should be made on the basis of maximum validity. When inference is made to a categorical criterion measure, however, decisions can be made or advice given with high confidence about individuals who have extreme scores on the test. Use of a well-designed sequential strategy is acceptable, and the injunction concerning use of short tests is modified accordingly.

Although we have not improved academic prediction much in the last 50 years, my first aspect of fairness is relatively clear-cut with respect to the research and statistical operations required. In contrast, the second aspect involves a host of problems. In the first place, any given individual belongs to a very large number of different demographic groups. Which ones should be selected for research? Or must we investigate all of them? If each demographic variable is divided into only two categories, and if there are as few as 10 such variables, 1,024 separate groups are defined. I have some confidence that groups showing the largest test differences are most likely to show intercept differences with the consequent constant error in the estimate of the criterion score for any given individual, but this is far from being an infallible guide. As a matter of fact, it is a far more accurate guide to the sources of discontent and of political pressure to discontinue the use of tests than it is to the dimensions of test accuracy and the related aspects of fairness of inferences drawn from tests.

A second source of problems with the constant error dimension is the number of examinees in the two or more groups being compared. Given a large enough N, any two groups will probably show a significant difference in intercepts. This is based on the premise that there are no completely zero differences in nature. There is a third source of difficulty also. If for a large N a single regression line provides, by chance, a very close fit to the separate regression lines for two or more groups differing in mean score on the test, a change in the reliability of the test or the addition of a similar test in a composite (Linn and Werts, 1971) will produce a difference in intercepts.

It appears obvious from the above discussion that we cannot use

tests with perfect fairness as herein defined. We simply do not have enough information, and the information required is all but unobtainable. Nevertheless, there is a limited amount of information available from which we can try to assess the degree of unfairness that has resulted from the common practice of using a single regression equation for all demographic groups. There is no reason to be distressed by the lack of a perfect fit between any mathematical model and the real world—such discrepancies are ubiquitous, not rare—as long as the lack of fit is sufficiently small.

The problems of male-female and black-white differences have been most extensively studied. There are, in addition, studies of social-class differences and at least one adequate study of regional differences. Reports on Spanish-speaking Americans and American Indians are scattered and inadequate.

First, it is well to define what I mean by adequate research. If there is an independent continuous criterion measure, the research must involve a comparison of the regression equations. The two characteristics of these equations that are critical are, as noted earlier, the slopes and the intercepts. (The preliminary assumption of equality of standard errors of estimate can, within limits, be violated. Inequality of standard errors of estimate that results from unequal slopes is the important phenomenon.) Unfortunately, the literature on test fairness has become confused—one might even say contaminated—by reports of research incorrectly analyzed. Inability to reject the statistical hypothesis of zero slope in a small sample of a minority group does not constitute evidence for differential validity of the test for that group.

When the literature reporting regression comparisons is summarized, the following conclusion seems warranted: there is relatively little difference in the slopes or intercepts of regression lines as a function of the demographic groups that have been studied. Use of a single regression equation for these groups leads to no substantial degree of unfairness in drawing inferences concerning the criteria measured. The amount of error involved is generally less than the sampling errors of regression coefficients based upon Ns of the size typically found in validation studies for minority groups.

If needed, this conclusion can be made more detailed. The small errors resulting from the use of a single equation are of some theoretical importance. Slopes of lines for males and females are about equal, but there is a fairly consistent tendency for female performance on the criterion to be underpredicted by a small amount, that is, women tend

to perform at least as well as their test scores indicate. Slopes of regression lines for black students, particularly for black males, when compared to white students may be a little lower. There is also a fairly consistent tendency for the constant error for blacks to be a small degree of overprediction; that is, they tend to perform at best at the level indicated by their test scores.

As hypotheses to explain these discrepancies I suggest the following: a higher mean motivational, or other important noncognitive, level for females, and a higher contribution to criterion variance of motivational or other important noncognitive components for black males.

Note that the above conclusions are not invariant laws of nature. If there is sufficient intervention in training, or if the criterion measures differ for the groups compared (Bowers, 1970), a single regression equation is no longer a good approximation for both blacks and whites. It should also be noted that regression comparisons have not been made for criteria in early childhood education and have not been made for criteria removed in time by several years from the point of test administration. Regression comparisons have been made, however, for appropriate criteria in education, industry, and the military with comparable results in all of these areas. All in all, there are so few exceptions to the preceding conclusions that I feel no great concern about the unfairness of inferences from tests when a single regression equation is used to predict practical criteria.

The lack of empirical support for big differences in either intercepts or slopes of regression lines should not be surprising. There was never any good theoretical expectation that such differences would be found. The assumption of cultural deprivation does not in itself constitute a theory. When one starts with that as a premise and develops the intervening steps in the reasoning in order to form a theory, the expectation is that deprivation will depress the scores on both the predictor test and the criterion measure, but not the correlation between the two. Approximate identity of regressions for minority and majority groups is the result anticipated.

The only theoretical bases for a different expectation that have occurred to me will be instantly, and correctly, found unacceptable by almost everyone in this audience. One is applicable only to racial comparisons and goes as follows: Negroes and Caucasians either belong to different species or the environmental differences have been so profound that the same psychological principles do not apply to both groups. Therefore, the organization of ability measures will differ

radically in the two groups. The second is hardly more plausible, even granting an antecedent in Spearman's "mental energy" concept of 50-70 years ago. In this second "theory," intelligence and other aptitudes are considered to be potential forces or mental powers like water under pressure or like stored electrical energy. Then the hydrant or switch is finally found and uncovered, after being buried by years of deprivation; intellectual power is available, fully developed, the minute the spigot is turned or the switch thrown.

I do not know of any careful studies of the content validity of achievement tests for different demographic groups. The problem has become confused by dragging in the red herring of discrimination and early deprivation. There is no reason to believe that standardized achievement tests, selected by a school to fit its curriculum, are not fair measures, as I have defined the term, for different demographic groups. A high school senior who reads at the sixth grade level may have had his level of achievement depressed by early deprivation, but the fact remains that he does not read very well. If reading is important, as it certainly is, both in further academic work and in citizenship, it is important to know the size of the deficit without regard to its possible causes.

In contrast to inferences concerning practical criterion measures, inferences are frequently drawn concerning capacity to learn from aptitude and intelligence tests. In this interpretation, these tests are sharply distinguished from tests of what has been learned. These hypothetical capacities are typically interpreted as fixed or constant and are frequently assumed to be innate as well. Such inferences go far beyond the available evidence, or are even contradicted by good evidence, and *are* highly unfair to all examinees, both children and adults, and members of both privileged and underprivileged groups.

In the first place, if there are capacities to learn, they are not measured directly by any present psychological or educational test. The very act of testing requires that the examinee have acquired a repertoire of learned responses. Secondly, the functions measured by intelligence and aptitude tests are not stable in the individual over time. Intercorrelations among a series of tests of the same function administered over a period of time invariably fall into a simplex pattern. Test-retest correlations over a time span are never as high as the respective reliabilities, and continue to drop as time increases. Finally, there is the issue of innateness of the function. If we knew the correlation between phenotype and genotype, we would estimate a score for

the genotype and attach to it a standard error, but we do not know that correlation. In place of knowledge we have widely varying estimates of the correlation with the variance of the estimates depending more upon the assumptions of the person making the estimate than upon the size of his sampling error. If the correlation were known, however, estimation would proceed by regression methods in a fashion parallel to the estimation of true score on the test.

There is some evidence that the size of the correlation between phenotype and genotype, although the level is uncertain, differs as a function of social class (Scarr-Salapatek, 1971). It is not at all improbable that this correlation varies more from one demographic group to another than do the correlations between test scores and practical criteria. For the present, at least, a test score has more meaning and can be used more fairly in making behavioral predictions than in making inferences concerning theoretical constructs. This may even be true for the relatively modest construct of true score.

Up to this point I have avoided the concept of the fairness of the use of tests for selection purposes, concentrating instead on the fairness of the inferences about individuals. Does this distinction have meaning? From the growing literature in this field, one can conclude that many persons would agree that it does.

Thorndike (1971) has contributed a contrast of two definitions of fairness, and Darlington (1971) has furnished a broader discussion. Darlington's exposition has recently been expanded by Cole (1972). Darlington presents three correlational definitions of fairness in selection in which value judgments are somewhat hidden, a fourth which requires an explicit value judgment for each criterion, and a fifth which is defined without reference to criteria. I shall discuss for the moment only the first three and shall neglect the superficial fifth entirely.

If the same qualifying score is to be used for members of both groups, the following relationships are required if the test's use is to be considered fair. In these three definitions of fairness x represents the test, y the criterion, and z the demographic group (coded 1 for majority, 0 for minority).

A) $r_{xz} = \frac{r_{xz}}{r_{xy}}$, derived from $r_{xz \cdot x} = .00$

B) $r_{xz} = r_{xz}$, which equates the probability of meeting minimum criterion proficiency with the probability of qualifying on the test

C) $r_{xz} = r_{xz}r_{xy}$, derived from $r_{xz \cdot y} = .00$

Definitions A and C above require that the slopes of the regression lines for test and criterion in the two groups be essentially identical. If this assumption has to be discarded, different qualifying scores are necessary under definitions A and C; if the equalities have to be discarded, different qualifying scores are necessary under all three. When different qualifying scores are required, they are computed in accordance with the logic incorporated in the several equalities.

These definitions have very different properties and lead to very different consequences which both Darlington and Cole have discussed. I shall describe a few properties, but before proceeding I shall translate Darlington's definitions into a slightly different form. This is done in Figure 1.

Perhaps the first property to note is that Definition B represents the limit of the correction for attenuation of the regressions portrayed in A and C, but B itself is independent of the correlation between x and y. Secondly, there is no constant error in the prediction of criterion score from test in A, there is overprediction for the minority group in B, and still larger overprediction in C. However, this situation is exactly reversed for the regression of test score on criterion with no constant error for C, some for B, and still more for A. Definition C will qualify most minority group members, A least, and B is intermediate, but if level of performance on the criterion is important, or if elimination rate is a problem, A is preferred over B and the latter over C.

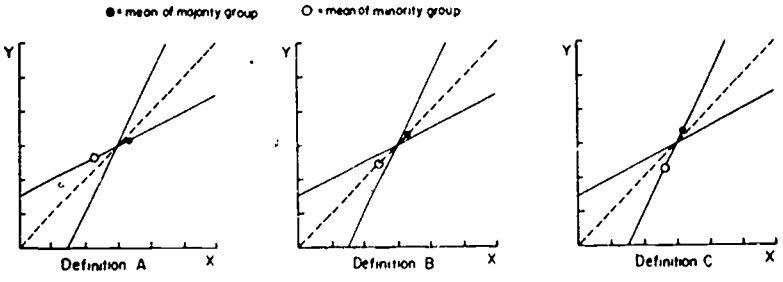


Fig. 1. Regression illustrations of three definitions of fairness for the use for a single qualifying score in selection for majority and minority groups.

One dimension of difference among these definitions not discussed by either Darlington or Cole is the extent to which selection fairness as here defined can be sabotaged by either the selecting institution or the examinees. A lesser slope of the regression lines for the criterion and test for the minority group will qualify fewer minority members under A and more under C. An institution might be motivated, and would be able, to reduce validity under A, while the examinees might be motivated, and would be able, to reduce validity under C. Definition B puts pressure on both groups to maximize validity. However, policing of institutions to prevent this type of unfairness is much easier than policing examinees so that the disadvantage of A is minimized.

Which of these different definitions of fairness is fairest? Cole presents C in an intuitively appealing manner, but intuitive appeal is a function of manner of presentation. When shown in regression form, the lack of bias in predicting test score from criterion measure, which is characteristic of C, clearly lacks appeal. For what purpose is prediction of test score desirable? Definition B has more properties that appeal to me than does C, although it suffers from an ailment common to all three: namely, the effects of measurement error. Measurement error in the predictor is critical for A, in the criterion measure for C, and a differential amount in test and criterion for B. Criterion measures, it should be noted, are typically less reliable than predictors. Definition A is, of course, the model for fairness for an individual, and the absence of sizeable constant errors in its use make it an attractive candidate for the preferred definition of fairness in selection as well.

For the present, however, I am most favorably inclined toward Darlington's fourth definition which involves making a value judgment for each criterion concerning the size of a bonus on the criterion for a particular minority group that is warranted by social and political considerations. Errors in prediction of both types are weighed against each other for each demographic group in determining the size of the bonus, and the logic of Definition A is used in selection but with the regression between predictor and modified criterion replacing the one ordinarily used.

Though I support Darlington's definition in which explicit value judgments are required, I do so reluctantly and with the hope that the need for a definition of this sort will pass. The present acute social problem which is the basis for the need was brought on by our failure to treat and to evaluate each human being as an individual. It is my hope that we can shortly return to that ideal and make it work. In-

dividual differences are very real and very important. In a recent personal letter to me, Professor E. R. Hilgard summarized this matter about as follows: if individual differences did not exist, we would have to assign them.

It is possible, even probable, that the problem of constant errors as a function of group membership under Definition A would virtually disappear if we were only better psychologists. Linear composites of measures covering a wider gamut of important human characteristics than are now being measured could reduce present constant errors in prediction substantially. In contrast, increasing the reliability of present predictors would have very small effects. I am willing on the basis of a little data and more theory to substitute "would" for "could" in the statement concerning the effects of linear composites and thus convert a true statistical statement into a psychological hypothesis. This hypothesis says that if we were able to predict more accurately for individuals in the majority group, we would need to be concerned less with the constant errors associated with group membership when predicting across groups. This also means that a good deal of research in recent years has been misguided. In place of looking for uniquely black or white, male or female, or lower-class, middle-class predictors, we should have been looking for better human predictors.

There is still another possible definition of fairness in selection that, to my knowledge, has not yet been suggested. This is to transform the selection problem into the classification problem and use the multiple discriminant function in place of multiple regression. The late Philip Rulon convinced me many years ago (see Humphreys, 1952) that discriminant thinking had a great deal of merit. Neither he nor I convinced a substantial number of others, however, and the use of the technique has not grown as I hoped and expected that it would.

Note that Definitions B and C partake of the logic of the discriminant in that both assume that meeting minimum standards on the criterion is sufficient: that is, an examinee either belongs in the criterion group or he does not. It might also be noted that college admissions officers utilize similar thinking, particularly with respect to the noncognitive traits of entering students. They believe that a particular type of student will be best satisfied and will fit best in their particular institution. In many, many situations we could select in both industry and education on the basis of those traits that are found in currently successful, satisfied workers and students. Instead, each institution now tries to maximize the aptitude level of its selectees. More does not

always mean better. For those who would like to look further at the mathematics of the discriminant problem, Tatsuoka's new book (1971) is recommended. Incidentally, the book is dedicated to Philip J. Rulon.

In closing it is legitimate to inquire concerning the reasons for the attacks on psychological and educational tests. In part, tests are all too frequently given under poorly controlled and unstandardized conditions. In patt, also, the attaeks are the result of the unfair inferences concerning native capacities that have been described. In part, however, the attaeks are misdirected. A good test carefully administered furnishes good information which can be interpreted without regard to the demographic membership of the examinee. Attacks that assume the reverse of the preceding statement not only cannot be documented by good research, but what is worse they misdirect attention from the primary problem; namely, a large number of children are not learning to read or do arithmetic or know science and social studies very well. Furthermore, while membership in any demographic group does not constitute insurance against inadequate learning, a large proportion of these children are found among a small number of ethnic groups. Thus, a problem which is basically an individual differences problem psychologically and biologically becomes a problem in group differences with important social and political consequences.

Slow learning cannot be overcome by abolishing the devices that reveal it most clearly Any experienced teacher can go into inner city classrooms and reach a conclusion similar to the one reached by inspection of test results. Ratings differ from tests in being a little less reliable, a little less valid, and a great deal more subject to constant error, but basically they tell the same story. I am forced to conclude that moves to abolish tests are more ostrich-like than human-like. The problem will simply not go away.

Although the primary problem is slow learning in the schools, and elsewhere, the primary causes for this do not appear to reside in the schools. This is clearly true for black children on whom we have the most and best data. Although these children are much below average in achievement at the end of the 12th grade, they also show an approximately equal relative deficit in the first grade. Thus the schools do not produce the deficit, but neither do they compensate for it.

If the schools are not responsible for the initial deficit, where do the causes lie? There are numerous possible sources, but no one source can be tagged with confidence as making a given precise contribution

to variance. There is good reason to believe, however, that the contribution of each is nonzero. The list starts with genetic differences and includes the totality of the environment from the development and release of the ova and sperm through fertilization, the prenatal, perinatal, and postnatal environment, to the entrance of the child in school. Environmental effects that are biological may be as irreversible as the genetic constitution, and the effects of many early learnings are highly resistant to change.

When looked at in this way, a solution that simply envisions new buildings, more books, smaller classes, more advanced degrees for teachers, that is, more money for education, is inadequate. Just doing more of what has been done in the past is very unlikely to help appreciably. Within the limits of formal education the only hopes are radically new curricula and radical changes in instructional techniques. While there can be no guarantee that such changes will solve the problem, these are areas of much higher priority for research and development than are new intelligence and achievement tests.

Relevance of curriculum content to life experiences seems highly desirable as is the gearing of this content to where children are in their intellectual development rather than where we think they ought to be or where we would like them to be. For example, standard English should be the terminal goal for the public schools, but to reach this goal it is probably necessary to compromise in the early grades. With respect to instructional techniques that differ radically from present methods, my ordering of the degree of promise places self-paced, including computer-based, techniques teamed with adequate reinforcement provisions in first position. A change from conventional schools to the so-called free schools, however, would further depress already depressed academic performance.

I shall not try to suggest changes for the preschool conditions previously summarized. These problems, however, are the responsibility of both the larger society *and* of the communities and families in which children are conceived, develop, and learn.

Questions and Answers

Q: What do you consider a compromise to demanding English in the primary schools?

A: The use of Spanish by the teacher in the primary grades; the use of readers in ghetto English, but with the goal constantly in mind that we have standard English learned and used by the end of the public school period.

Q: What are the "free" schools that further depress academic performance?

A: By "free" schools I do *not* mean the well-organized English day schools or open classrooms. When well done, such classrooms come close to meeting my specifications for the preferred direction in education. I do refer to unorganized, laissez-faire schools that have appeared in this country. I have only a little data specific to these schools, but they do violate almost everything we know about learning.

Q: A question about my conclusion that "doing more of the same" is not going to be very helpful.

A: My choice of words was unfortunate, although the context should have given the clue. "Providing" is more accurate than "doing." More money for education is not the answer. I recommend the new book by Jencks and others concerning the influence of the schools on intellectual performance, occupation, income, and so on.

Q: Is there any way of taking into account the possibility of systematic bias in the measurement of criterion performance?

A: Criteria are of necessity culturally bound; that is, criterion performance is something valued by the society. It should be measured objectively if at all possible. Ratings used as criteria are subject to bias which is difficult to control.

Q: How did your research lead to your conclusion that greater accuracy for individuals would lead to smaller constant errors?

A: I have no research, but there are some important general principles. In the gene pools of blacks and whites there is almost 100 percent overlap in the genes stemming from a single evolutionary origin for for the races of man. There is also much, much more overlap in the environment for blacks and whites in this country than there are differences: schools, diet, TV, radio, language, newspapers, and the like. There is very great overlap on all psychological traits.

Q: Black-white differences have commonly been found to be minimal in grade one. How can you conclude that school differences are small?

A: The premise is not true. The mean difference between blacks and whites in this country in standard score units is about the same in the first grade as in the 12th. Mental age or grade equivalent units become smaller with advancing age and movement through the grades. A one-unit deficit at age six is the equivalent of a two-unit deficit at 12, a three-unit deficit at 18.

Q: Are speeded tests "fair"?

A: It depends on whether speed increases or decreases validity.

Q: What are the prospects for the new cortical response measures?

A: Nil for replicating the information we now obtain from an intelligence test; but something useful may develop.

Q: Are there studies of interrelations of abilities as a function of demographic data?

A: Yes. Such studies date back to the early fifties in the Air Force. There are minimal differences in the organization of abilities as between males and females, blacks and whites, or low and high SES groups.

Q: How do you exclude broader life experiences and focus only on commonality in curriculum experiences as important influences on the organism's development?

A: I do not, but one gives an achievement test in order to find out how well a student reads or does arithmetic. There are many possible explanations for a deficit, but explaining the deficit does not erase it.

Q: Would you favor adjusting aptitude test scores on the basis of the mid-parents' educational level?

A: Not at all. I favor maximizing the accuracy of predictions or inferences. As a general procedure this adjustment would decrease the accuracy of predictions. For certain predictive situations, this variable is useful, but it should be studied independently.

Q: Why do you use the term "slow learning"?

A: In part, I use it as synonymous with below average academic performance, but it also suggests a useful concept. If we consistently allowed time to vary and made certain that students acquired the skills and knowledge required before moving on to new material, some below average performers might achieve higher levels of competence than they now do.

Q: Is it true that intelligence tests *overpredict* performance for blacks, because performance is also predicted by *achievement*, and achievement is reduced relatively more by an underprivileged background than is intelligence?

A: There is no qualitative difference between achievement and intelligence tests. My hypothesis is that overprediction is the result of (a) errors of measurement and (b) failure to measure important functions.

REFERENCES

Bowers, J. The comparison of GPA regression equations for regularly admitted and disadvantaged freshmen at the University of Illinois. *Journal of Educational Measurement*, 1970, 7, 219-225.

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

Cole, N. S. Bias in selection. *ACT Research Report No. 51*, May 1972. Iowa City, Iowa: The American College Testing Program, 1972.

Darlington, R. B. Another look at "cultural fairness." *Journal of Educational Measurement*, 1971, 8, 71-82.

Gulliksen, H., & Wilks, S. S. Regression tests for several samples. *Psychometrika*, 1950, 15, 91-114.

Humphreys, L. G. Individual differences. *Annual Review of Psychology*, 1952, 3, 131-150.

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

Scarr-Salapatek, S. Race, social class, and I. Q. *Science*, 1971, **174**, 1285–1295.

Tatsuoka, M. M. *Multivariate analysis. Techniques for educational and psychological research*. New York: John Wiley, 1971.

Thorndike, R. L. Concepts of culture-fairness. *Journal of Educational Measurement*, 1971, **8**, 63–70.

A Theoretical Approach to Cultural and Biological Differences

CHARLES V. WILLIE
Syracuse University

I have entitled this presentation "A Theoretical Approach to Cultural and Biological Differences" because I intend to draw my inferences more from the similar findings of the scholars rather than setting forth some findings of my own. Lest the audience become concerned, however, that I am some kind of mutational freak, I will not behave completely out of character as a discussant. If the spirit should so move me, I will get in a few of my own substantive licks here and there and pretend that they are elaborations on a theme initiated by one of the authors.

Essentially, J. McV. Hunt, Lloyd Humphreys, Eleanor Maccoby and Carol Jacklin have focused upon assumed biological differences or differences thought to be contributed to by heritability more so than by environmental circumstances (Jensen, 1969, pp. 1–123). They have analyzed variations in such phenomena as intelligence, aptitude; achievement, verbal, mathematical, and spatial ability by sex, race, ethnic, and social-class categories. Hunt states that "heredity is clearly primary" with reference to intelligence. But Humphreys warns us that "the functions measured by intelligence and aptitude tests are not stable in the individual over time." And Maccoby and Jacklin concluded "It is still a reliable generalization that there are no sex differences on these [intelligence] tests." From these statements we may determine that populations with obviously different hereditary charac-

teristics such as those associated with sex respond similarly to intelligence tests, and that a person whose hereditary characteristics are obviously the same at birth as in later years responds differently to intelligence tests at different periods in time. If under these conditions, if heredity is clearly primary, the question must be raised: primary for what?

Ours is a discussion about population genetics. I thought it appropriate to consult an outstanding scientist in this field, Theodosius Dobzhansky. The second edition of his excellent book, *Genetics and the Origin of Species*, was awarded a prize by the National Academy of Sciences (Dobzhansky, 1951). From Professor Dobzhansky's discussion, I have extracted eleven principles which make specific contribution toward a theory of cultural and biological differences:

1. A Mendelian population is . . . a reproductive community of individuals who share in a common gene pool [p. 15]
2. Gene frequencies and variances, rather than averages, characterize Mendelian populations. All Mendelian populations are polymorphic [pp. 108–109].
3. A species [is] polymorphic if it contains a variety of genotypes, each of which is superior in adaptive value to the others in some habitats which occur regularly in the territory occupied by the species . . . [pp. 132–133]
4. Polymorphic populations [are], in general, more efficient in the exploitation of ecological opportunities of an environment than genetically uniform ones. . . [pp. 132–133].
5. Racial differences are more commonly due to variations in the relative frequencies of genes in different parts of the species population than to an absolute lack of genes in certain groups . . [p. 176].
6. Race and species are populations . . which remain distinct only so long as some cause limits their interbreeding [p. 18].
7. The sum of genes of an individual or a population constitute the genotype. . . The resulting bodily forms . . are different phenotypes. . . A genotype is potentially able to engender a multitude of phenotypes . . [pp. 20–21].
8. Any phenotype that may be formed is necessarily a response of the environment to the activity of the genotype. The genotype reproduces itself regardless of what phenotype it happens to evoke in a given instance [pp. 20–22]
9. Some genotypes permit a greater amplitude of modifications . . than others, and some traits are plastic while others are more rigidly fixed [p. 23].

10. Human intellectuality and emotional development is an example of a great plasticity and susceptibility to environmental influences. Environment, upbringing, schooling, association with other people and the manifold variations of individual biographies are powerful moulders of human personality. The genotypic determinants of human personality are easily obscured by the environmental ones [p. 23].
11. The populations of most species vary, often within an enormous range, from generation to generation [p. 163].

It is appropriate to close this review of findings pertaining to population genetics by returning to Dobzhansky's discussion of the Mendelian laws. According to Mendel, *the fundamental units of racial variability are populations and genes, not complexes of characters*. Dobzhansky goes on to say that:

> many studies of hybridization were made before Mendel, but they did not lead to the discovery of Mendel's laws In retrospect [reports Dobzhansky] we see clearly where the mistake lay: they treated as units the complexes of characteristics of individuals, races, and species and attempted to find rules governing inheritance of such complexes. [Dobzhansky states that] Mendel was the first to understand that . . . the inheritance of separate traits [and] not [the inheritance] of complexes of traits . . . had to be studied [Dobzhansky bemoaned the fact that] some of the students of racial variability consistently repeat the mistakes of Mendel's predecessors [p. 177]. [That is, they try to trace inheritance through complexes of characteristics.]

Dobzhansky concludes that "Race is not a static entity but a process. . . . Racial variability must be described in terms of the frequencies of individual genes . . . in groups of individuals occupying definite habitats. Such a description is more adequate than the usual method of finding the abstract average phenotypes of 'races' . . . [pp. 177–178]."

Please note that this review has focused on population genetics and not the genetics of individuals. As stated by Dobzhansky, the rules governing the genetic structure of individuals are different from those governing the genetic structure of a population Moreover, he states that "every human individual is unique, different from all others who live or lived [pp. 15, 4]."

Returning to the question posed earlier, I agree with Hunt that heredity is primary. It is primary for the continuation of the species; for only a genotype may reproduce a genotype. (As Hunt has said,

only an elephant can reproduce an elephant.) After that, much is left to the habitat and environment. On this point agree all of the authors as well as the population geneticists. For example, Hunt states that life under different circumstances can produce differences in a given genotype or a given population of genotypes. He makes specific reference to children from poor families who have lived under conditions of only poverty. Humphreys states that children who are much below average in achievement may suffer the deficit from a number of possible sources, starting with genetic differences and including the totality of the environment. Maccoby and Jacklin state that data on the incidence of specific deficits in learning abilities indicate that these occur considerably more frequently among boys than girls. However, they explain that the greater vulnerability of the male child to anomalies or prenatal development, birth injury, and childhood disease is well known and that this vulnerability probably does affect the incidence of very low scores on tests of intellectual abilities. These statements are similar to those of Dobzhansky, that a phenotype is a response of the environment to the activity of the genotype and that the genotypic determinants of human personality are easily obscured by the environmental ones. Thus, the findings of Hunt, Humphreys, and Maccoby and Jacklin are in accord with a fine tradition of behavioral science theory which, in summary, states that a genotype is potentially able to engender a multitude of phenotypes. Apparently most of our tests have been measuring genotypic responses to the environment, or phenotypes, which accounts for the instability of such measurements on the same individual at different periods in time.

Dobzhansky's finding that "human intellectuality . . . is an example of a great plasticity and susceptibility to environmental influences" is confirmed by the investigations of our authors and those of other scientists on whom they report. Hunt has shared with us the results of studies conducted by him and associates in the Parent and Child Center in Mt. Carmel, Illinois. He and his colleagues found that child-rearing of parents from a lower class has been improved by a parent education program so that the behavior of their children in the development of object permanence surpasses that of the middle-class, at least during the first two years of infancy. These and other findings are illustrative of the great plasticity of intellectuality. Humphreys states that there is some evidence that the size of the correlation between phenotype and genotype differs as a function of social class. This evidence may or may not indicate genotypic plasticity; but it

certainly does point toward phenotypic variability, possibly related to differential environmental circumstances of life.

Maccoby and Jacklin inform us that variations in phenotypes, particularly with reference to spatial and verbal ability, may occur at different age levels because of different contemporary and past experiences. While they hold strongly to the general statement that there appears to be no basis for saying there is a sex difference in overall intellectual power, mention is made of variations in the stages in which these specific skills develop. For example, an increasing superiority of girls on verbal tasks and an increasing superiority of boys in spatial tasks and possibly some quantitative tasks have been observed. Thus far, Maccoby and Jacklin have not located any solid research which might explain social factors associated with these different patterns of development. However, it is important to point out the need for longitudinal studies on patterns of sequential development. While the females may have a head start in verbal development during adolescence, presumably the males play catch-up and do make creative contributions in arts and letters in adulthood. Indeed, it is my guess that the honorary degree, Doctor of Humane Letters, is given more frequently to men than to women during the May-June college commencement pageant. If this be so, again it is probably due not to any inborn differences in intelligence between the sexes but to the response of male genotypes to a sexist environment. Be that as it may, the fact of the possibility of delayed development is worthy of mention for behavioral scientists who have forgotten that adaptation to the environment is one of the chief means of survival.

At this point may I introduce a few findings from one of my studies of black students at predominantly white colleges. These findings have to do with their academic adaptation. As you might guess, most black students enrolled in the four upstate New York colleges which I studied in 1969-1970 did poorer than most white students. In terms of self-reported cumulative average grades, about 23 percent of the blacks compared to 49 percent of the whites had As or Bs; 64 percent of the blacks compared to 49 percent of the whites had cumulative grade averages at the C level, and 13 percent of the blacks compared to 2 percent of the whites had self-reported cumulative grades of D or less. Averages like these tend to mask so much and contribute to our misunderstanding about variability in adaptation. When analyzed by year in school, my colleague and I discovered that black college seniors had better grades than black college freshmen. But not only

that; black college seniors had better grades than white college seniors: 52 percent of the black college seniors compared to 42 percent of the white college seniors had cumulative grades at the A and B levels (Willie and Sakuma, 1972, p. 86).

Do these findings mean that black college seniors come from intellectually more gifted parents than black college freshmen? I doubt it. Do these findings mean that black college seniors come from genetic pools that are intellectually superior to the genetic pools from which white college seniors are drawn? I doubt this possibility too. My conclusion would be that the superior achievement of these black seniors compared with other black and white students is a function of the way in which they adapted to the difficult situation in which they found themselves. Through endurance, senior black students eventually transcended and overcame many academic obstacles. The superior outcome of their endurance did not become visible until the fourth year. During the first three college years, the average black student trailed behind the average white student in academic achievement as measured by grades. Many black students were unable to persevere until the fourth year. Still the number of black students in the freshman year in these upstate New York colleges is twice as large as the number in the senior year, while the number of whites in each of the four classes is more evenly divided. It is probably fair to say that both the low achievement of black students during the first three years and high achievement of blacks during the fourth and last year were due to environmental circumstances and adaptations thereto. Also, as Humphreys has pointed out, motivational components come into play.

On the basis of my study (which unfortunately is inconclusive because it is limited to cross-sectional data), I would assert that black students in predominantly white colleges are neither superior nor inferior to their white college mates, that each has an overlapping range of intellectual capacity which is capable of making a variety of responses to different environmental situations, and that black seniors tend to respond by superior academic performance while black freshmen, sophomores, and juniors respond by inferior academic performance compared to whites. We can understand this only if we can remember four principles earlier set forth by Dobzhansky: (1) that "a genotype is potentially able to engender a multitude of phenotypes," including those which function in superior and inferior ways, (2) that "genotypic determinants of human personality are easily

obscured by the environmental ones," (3) that "racial differences are more commonly due to variations in the relative frequencies of genes in different parts of the species population than to an absolute lack of genes in certain groups," and (4) that "environment, upbringing, schooling, association with other people and the manifold variations of individual biographies are powerful moulders of human personality."

Hunt and Humphreys have cautioned against treating measures of phenotypes—that is, measures of developmental achievement—as if they were genotypic. But I am a mite unhappy at the timid way in which all three papers dealt with our common tendency to rely upon statistical measures of aggregated characteristics to get at the problem of biological difference (if any) between the races, particularly as it relates to intelligence. Even though, as a racial group, their cumulative average grade attainment was lower, black senior college students in my study performed better than whites academically. They could not have performed better than whites academically the last year of college if they had not had the capacity to do so, a capacity which was probably present during the earlier three years.

It would appear that I am making a tautological statement that black students outperformed whites because black students had the capacity to perform as well or better than whites. This seemingly tautological statement, however, is of value because we know that some black populations perform less well compared to whites. Nevertheless, we should be reminded that even in such populations of poor performers the capacity to perform as well probably is present and that it could become manifest, given the appropriate set of circumstances or motivational components. One reason for not recognizing this fact is our tendency to rely either on composite measures of intelligence or on composite descriptions of a population. And thus we commit the same error committed by Mendel's predecessors, that of treating as units the complexes of characteristics of races or intelligence rather than recognizing as did Mendel that what must be understood is the inheritance of separate traits and not the inheritance of complexes of traits. Moreover, it should be stated again and again, that genotypic traits may be present even though they are not observable in the phenotype. This simple fact is frequently forgotten. In a gentle way, I am trying to say that although our statistical methods and techniques for studying variations (if any) in the association between intelligence and race appear to be sophisticated, conceptually—

especially with reference to the nature and form of heritability—some are pre-Mendelian and, therefore, dated.

Mention of the need to disaggregate traits for the purpose of studying heritability leads directly into a discussion of race. It is time we ceased the silly business of discussing variations in behavior by racial categories as if the races of humankind which we commonly recognize were pure. Indeed, racial purity would be a liability. Such a population would be less adaptable and less able to exploit its environment. Populations of Negroid and Caucasoid people in the United States, for example, are imaginary and at best abstract statistical constructs. Dobzhansky tells us that the genetic structures of populations can be molded into new shapes through the influences of selection, migration, and geographic isolation, and especially the breeding of species (Dobzhansky, 1951, p. 15). To be sure, there have been laws against the intermingling of the races in this country. But historian John Hope Franklin tells us, "the slave woman was frequently forced into cohabitation and pregnancy by . . . her master." He describes the miscegenation which went on during the slave period as "extensive (Franklin, 1967, p. 204)." Moreover, he indicates that there are records of marriages of Negro-White couples and Negro-Indian couples in New England during the colonial period (Franklin, 1967, p. 109). We know that there was considerable interbreeding between whites and the Native Americans also known as American Indians when the West was settled (Brown, 1970). In summary, there has been a lot of race mixing and interbreeding in the United States. The diversity of inherited characteristics exhibited by the people in any public gathering is ample evidence of the extensive cohabitation between all sorts and conditions of people over the years in this land. Lloyd Humphreys is right when he states that any given individual belongs to a very large number of different demographic groups. It is inappropriate to measure intelligence as a complex of characteristics, if one wants to understand something about inheritance. Maccoby's and Jacklin's approach of looking at specific skills as well as the call by Humphreys for better human predictors are more promising than the search for an inheritable composite. Also it is inappropriate to relate a faulty measure of intellectual heritability which Humphreys calls a "hypothetical capacity" to race, which Dobzhansky calls an abstract statistical phenotype, if one wishes to understand the association between innate characteristics. Neither intelligence as presently measured nor race as presently defined are innate. Yet we persist in correlating the two and

thereby compound our error by making what Humphreys would classify as "unfair inferences about native capacities." The discussion about race and intelligence in the United States, then, is so much talk about nothing. Measures of intelligence are unsatisfactory and so are the definitions of race. So what is all the fuss about?

I am inclined to believe that the controversy has little, if anything, to do with science. It seems to me that the controversy is a continuation of the social Darwinism in American thought so excellently documented by Richard Hofstadter (Hofstadter, 1955). His chapter on "Racism and Imperialism" details how Americans rationalized oppression of outgroups in the past as a natural development in which "backward races would disappear before the advance of higher civilizations [p. 171]."

Herbert Spencer, William Graham Sumner, and nineteenth century white Americans may have believed these thoughts. They used the findings of population genetics as a way of putting people down as inferiors and explaining outgroup failures. But twentieth century Americans have been exposed to more enlightening thoughts. They know "that the physical well-being of [human kind] is a result of their social organization and not vice versa." They know that "social improvement is a product of advances in technology and social organization, not of breeding or selective elimination (Hofstadter, 1955, p. 204)."

If twentieth-century Americans know these things, how do they continue to use social Darwinism? It seems to me, in the light of the discussion above and some unpublished data which I have on a school desegregation study, that social Darwinism is used today not so much to put down the outgroup as subhuman as to build up the ingroup as superhuman. Also social Darwinism now is used to explain the lack of success of the ingroup rather than the failure of the outgroup.

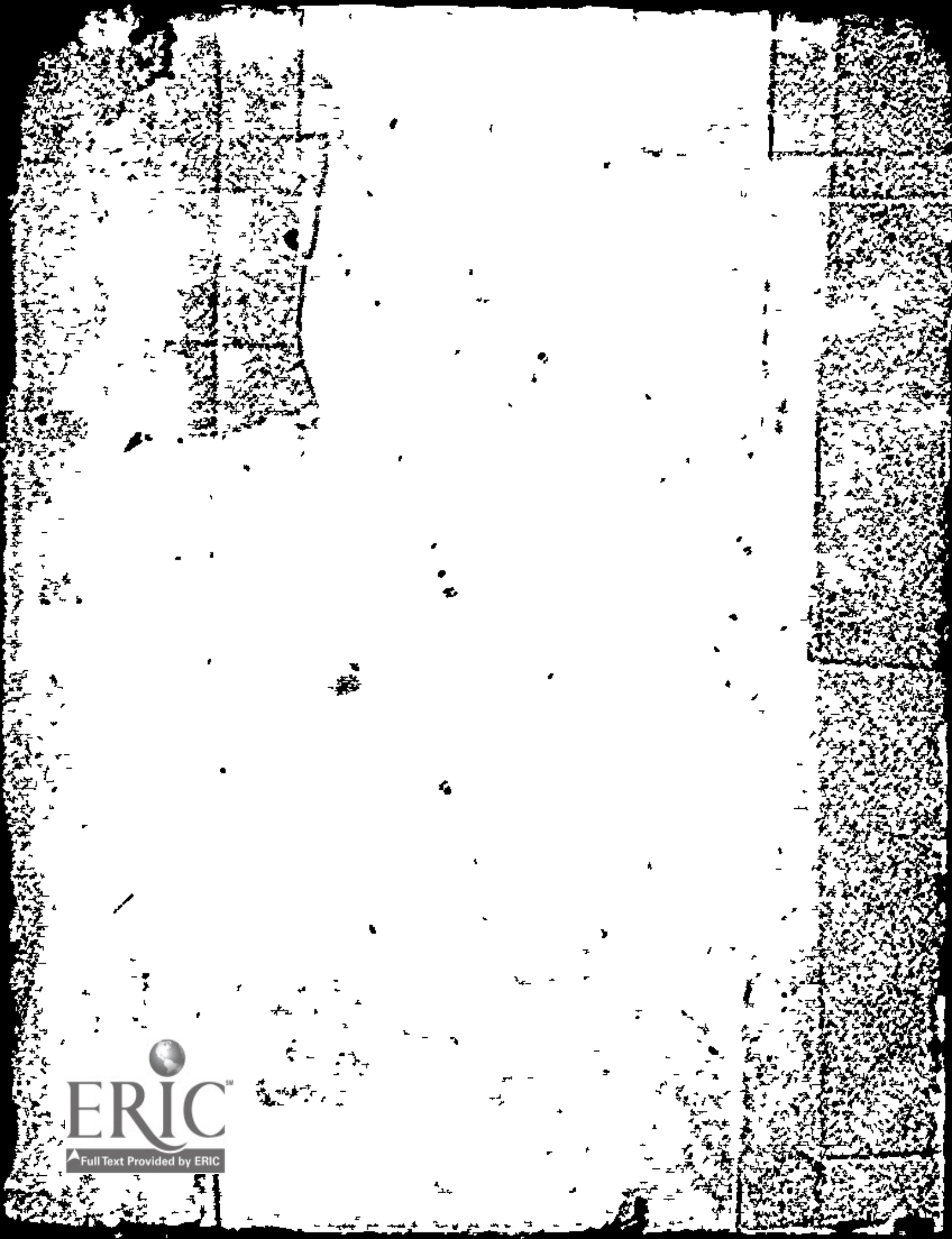
In a few predominantly white elementary schools in upstate New York which had recently received a modest number of 50 to 60 black children, white teachers were asked to rate the level of social adjustment for each new child in their class. New children were inner-city blacks transferred to schools in middle-class communities to improve their racial balance and affluent whites who were new residents in the neighborhoods surrounding these schools. Social adjustment was rated on a multi-interval scale ranging from well-adjusted, fairly well-adjusted, moderately adjusted, to poorly adjusted. Children also were requested to rate their degree of adjustment on a multi-interval scale.

With reference to the well-adjusted, white teacher ratings of black children tended to correspond closely with the ratings which black children gave themselves in this area. For white children, however, the proportion whom white teachers considered to be well-adjusted substantially exceeded the proportion of the children who felt that way. The white teachers tended to see well-adjusted black children as they were; but they perceived the adjustment experience of several white children to be better than the white children believed it was.

While the literature continues to grow with accusations that white teachers are unmindful of the capacities of black children, I am inclined to believe that twentieth-century whites who have lived through the civil rights revolution of the 1950s and the 1960s in this country are not so unmindful today. They have no need to put down blacks or members of the outgroup. But because of the remnant of social Darwinism in American thought, there is now a tendency to inflate the capacity of the ingroup.

The article by Arthur Jensen on IQ and scholastic achievement is a classic example of the use of social Darwinism to explain away the lack of success of the ingroup (Jensen 1969, pp. 1–123). One does not have to read far into that article to pick up a twentieth-century tone of Manifest Destiny. Read this: "The remedy deemed logical for children who would do poorly in school is to boost their IQs up to where they can perform like the majority. . . ." This is a direct quote from Arthur Jensen. He goes on to say: ". . . this is in fact essentially what we are attempting in our special programs of preschool enrichment on compensatory education [p. 3]." He develops a series of questions. "Why has there been such uniform failure of compensatory programs whenever they have been tried? What has gone wrong? In other fields, when bridges do not stand, when aircraft do not fly, when machines do not work, when treatments do not cure, *despite all conscientious effort on the part of many persons to make them do so*, one begins to question the basic assumptions, principles, theories and hypotheses that guide one's effort. Is it time to follow suit in education [p. 3; italics added]?" He asserts that the success of preschool and compensatory programs is to develop gains in IQ and in scholastic achievement. And then, in a bold attempt to stake out the public definition of the problem, he states that "our diagnosis should begin . . . with the concept of the IQ."

First question. By what line of reasoning did Jensen determine that the intellectual adaptation of the majority should be the determinant



norm for the kinds of adaptations which the minority ought to make? Remember my upstate New York study of black college students. Minority black seniors functioned better than majority white seniors. It is an arrogant act to establish the ingroup as a model for all others to follow. Jensen's position with reference to black youth was quite similar to one advocated by Daniel Patrick Moynihan for black adults. Moynihan advocated overhauling the black family because he declared, "it is so out of line with the rest of the American society (Moynihan, 1966, p. 29)."

Again the majority is used as a model for the minority. When this is done, the majority may expect the minority to ignore it for the minority has found survival value in the kinds of adaptations it has grown accustomed to. Moreover, minority adaptations are sometimes more beneficial for all.

An honest assessment by majority and minority members of this society would conclude that "all conscientious effort" has not been expended to provide compensatory educational programs. A conscientious effort would require that a disproportionate amount of community educational resources be devoted to educating the disadvantaged. This never has been done in this nation.

Finally, the poor and disadvantaged have some ideas of their own about what they would like to get out of formal education. Manipulation of their children's IQ may not be their highest priority. Learning how to endure (Douglass, 1962, p. 39), and how to develop a positive concept of the self (Rosenberg and Simmons, 1972, pp. 21–30), and how to gain a measure of control over one's environment (Teele, 1970, p. 367) probably are as important to the poor as gains in IQ.

And so the great experiment failed which was fashioned by the ingroup for the outgroup. The affluent, majority, ingroup tried to make over the poor, minority, outgroup in its own image. Rather than accept the failure as an inadequately planned program, improperly imposed upon a population, the ingroup with Arthur Jensen as philosopher-king has turned once more to social Darwinism, this time to explain the lack of success of the ingroup. Higher Horizons, Head Start—they were misguided efforts to remedy the irremediable and not failures in social organization, according to twentieth-century social Darwinism.

The time has come to deal with cultural and biological differences as they ought to be dealt with, in the tradition of the social and behavioral sciences, and not in the tradition of social Darwinism.

Questions and Answers

The kinds of questions addressed to me are sociologically quite significant. I received six questions and five dealt with one issue—the issue of explaining away why black college seniors had better grades than white college seniors. Let me share those questions with you very quickly.

Q: Could one explain the high achievement levels of seniors as compared to freshmen on the basis, in part, of differences in the variations in the variances of the subgroups; that is, those who failed have already dropped out? Also, in comparing black with white seniors, one might be able to offer an explanation based on the view that those blacks who stayed were more highly motivated to stay than white students who stayed. The higher-scoring blacks were motivated by self and therefore this group of blacks could be better compared to the upper portion of whites.

Q: What controls are used in your study for differences in grades for black and nonblack students resulting from differences in courses? Were, for example, student-teaching grades considered equal to grades in physics, math, and chemistry courses?

Q: What percent of the entering black student group persevered to the senior year compared to the white student group? Did a larger rate of attrition among the former leave a smaller but better adapted group than the latter group represented?

Q: Are the blacks at Syracuse of middle or lower socioeconomic status? Perhaps there is an operative embedded.

Q: Aren't your upstate New York studies vulnerable on two points: differences in grading methods and selective factors?

A: Obviously, I cannot answer all of these questions in the limited time available. As a social scientist, I pointed out that my studies were inconclusive because they were based on cross-sectional data. But, I would say this: I believe my findings.

My major point, however, was not about the population of black students who were seniors. My major point had to do with individual black students who were seniors. The fact that they performed well as seniors meant that they had the capacity to perform well when they

were freshmen, even though their freshmen grades may or may not have reflected that capacity.

REFERENCES

Brown, D. *Bury my heart at Wounded Knee*. New York: Bantam Books, 1970.

Dobzhansky, T. *Genetics and the origin of species*. New York: Columbia University Press, 1951.

Douglass, F. *Life and times of Frederick Douglass*. New York: Macmillan, 1962.

Franklin, J. H. *From slavery to freedom*. New York: Knopf, 1967.

Hofstadter, R. *Social Darwinism in American thought*, Boston: Beacon, 1955.

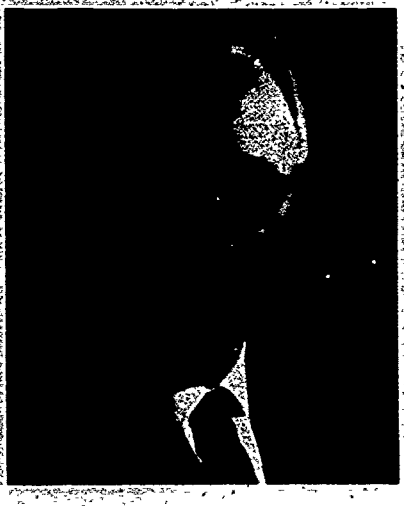
Jensen, A. R. How much can we boost IQ and scholastic achievement? *Environment, heredity, and intelligence*. Cambridge: *Harvard Educational Review*, 1969, Reprint Series No. 2, 1–123.

Rosenberg, M., & Simmons, R. G. *Black and white self-esteem: The urban school child*. Washington, D. C.: The American Sociological Association, 1972.

Teele, J. E. Sociocultural factors relating to mild mental retardation. *Social Science and Medicine*, 1970, 3, 363–369.

U. S. Department of Labor. *The Negro family*. Washington, D. C.: U. S. Government Printing Office, 1966.

Willie, C. V., & Sakuma, A. (McCord). *Black students at white colleges*. New York: Praeger, 1972.



Luncheon Address

Recycling the Problems in Testing

HENRY S. DYER

After accepting Bill Turnbull's pleasant but overwhelming invitation to deliver this luncheon address, I felt the need somehow to pull myself together, so I tried to catch my breath by doing a bit of casual research on the luncheon addresses of the past. One finding that rather shook me has to do with the first time a luncheon address got itself insinuated into the conference program. It was the year 1951. Imagine my surprise when I discovered that the conference chairman that year was none other than one Henry Dyer.

Another finding that our good friend Anna Dragositz dug up for me has to do with the ages of the luncheon speakers. She found that their ages at the time of speaking have ranged from 37 to 73 years with the median at 57 years. This puts me comfortably above the norm at the 79th percentile. Having no criterion reference for that score, however, I'm not sure whether it's good or bad.

A third finding was that out of the 20 addresses thus far delivered, only six have had much of a connection with testing problems per se. Therefore, all things considered, and because I think testing problems are as important as any problems in the world, I thought it would not be out of place to express a few thoughts about some of the old problems in testing as they appear in 1972.

Hence, the ambiguous title of this speech: "Recycling the Problems in Testing." The ambiguity was put there on purpose to give me some

room for maneuver on reaching this moment of truth. Since then, I have narrowed down the subject by lengthening the title. It should now read: "Recycling the Problems of *Educational* Testing with Special Reference to the Problematical Uses of Achievement Tests as Measures of Instructional Effectiveness in Large-Scale Testing Programs Involving Diverse Groups of People." (I shall be curious to see what the editor of the conference proceedings does with that mouthful.)

I

In the introductory chapter with which Anne Anastasi starts off her admirable selection of Invitational Conference papers (Anastasi, 1966), she gives a good deal of attention to a testing problem that was very much on the minds of the tiny band of testing leaders who came to the first three or four conferences back in the 1930s, presumably to lick their wounds. They were deeply worried about the many and ingenious ways in which the people in the schools were mishandling tests and misusing test scores—especially the tests and test scores then being generated by state testing programs.

These days state testing programs are being transformed into state *assessment* programs with a concomitant shift in emphasis from the guidance of students to the evaluation of schools and their educational programs (Dragositz, 1971). The old problems in testing, however, have survived the transformation. They are still around; only they have taken on some new and rather more scary dimensions. The reason should be fairly obvious to anyone who has been keeping up with the educational assessment and accountability acts that state legislatures, following the lead of the Congress since 1965, have been grinding out over the past few years. These acts reflect an explosion of public interest in the workings of the schools that may be historically unique. As a consequence, the field of education has become strewn with politics, and educational testing has become an instrument, if not a weapon, in the political process (Kirst & Mosher, 1969; Cohen, 1970). And this means that our worries today about the mishandling of tests and the misuse of test scores must embrace not only school personnel, but also politicians and the diverse and pluralistic constituencies they serve.

Accordingly, to do obeisance to the title of this talk by stretching a metaphor practically out of shape, one might say that the problems

in testing that were afflicting the schools in the 1930s have turned out to be nonbiodegradable and are therefore in need of a recycling job to prevent them from befouling the streams of education in the 1970s.

Please note. I did *not* say that educational tests themselves are inherently pollutants of the educational process, even though some of the more uptight critics of testing seem so to contend. To the contrary, most of us here, I suspect—and this emphatically includes me—would earnestly subscribe to the proposition that testing, when rightly conceived and properly handled, is absolutely indispensable to the management and improvement of instruction all up and down the line, from the classroom, to the superintendent's office, to the school board, and even to the halls of the Congress. The trouble is that, as testing has become so much a part of our socio-educational culture, more people than ever before are unaware of the problems with which conscientious testers themselves have been perennially concerned—problems that must always be taken into account if testing in the schools is indeed to be rightly conceived and properly handled.

One disturbing manifestation of the difficulty is the tendency among the uninitiated to expect from tests harder information than the scores can reasonably be expected to supply. For example, a few years ago, when the contagion of performance contracting was in its incipient stage, I found myself in conversation with a government official who was one of the more enthusiastic proponents of the performance contracting idea. I tried to make the point that the gain scores on even the most reliable tests in reading or math or anything else were insufficiently devoid of measurement error to justify the exact payment by individual student results that was then being advocated. His reply was swift and off the point. It was high time, he said in effect, that test makers got on the ball and began producing tests that they could guarantee to be 100 percent reliable under all conditions.

Well, how does one penetrate the fantasy world exemplified by that kind of demand? How does one get across the shocking truth that 100 percent reliability in a test is a fiction that, in the nature of the case, is unrealizable? How does one convey the notion that the test reliability problem is not one of reducing measurement error to absolute zero, but of minimizing it as far as practicable and doing one's best to estimate whatever amount of error remains, so that one may act more cautiously and wisely in a world where all knowledge is approximate and not even death and taxes are any longer certain?

Take another example. Last year the education committee of one of

the state legislatures came up with an educational accountability bill that read in part as follows:

> If the performance of any school district on any test approved by the state board of education . . . does not equal or exceed the national performance average for such a test for two successive years, said school district shall not receive any further state financial assistance . . . until such time as said school district has achieved such national performance average (Kansas, 1971).

Somebody—one hopes it was someone in the testing fraternity—must have been sufficiently persuasive to convince that committee that the proposed legislation contained some fatal psychometric, not to mention educational, flaws, for the bill was not enacted into law. Nevertheless, it typifies some of the confusion in high places regarding the nature of testing and the proper uses of test scores. This confusion seems to be rooted in the almost irresistible tendency to hypostatize such an essentially meaningless abstraction as national performance average in tested achievement. The confusion is further confounded by the notion that a test norm can double in brass as a "standard" of performance to be reached by everybody. And, of course, attendant on these confusions is the amazing idea that the way to upgrade public education is to withhold funds from the schools whose students are not doing well.

In the old educational theology, it was generally supposed that the way to promote learning in the young was to beat the devil out of the pupils. The new educational theology seems to hold that the same state of grace can be more readily achieved by beating the devil out of the educators. There are some of us, I suspect, who would have doubts about the efficacy of either method of exorcism.

Now for a third example in a quite different domain. It comes from an earlier time, but it points to a number of testing problems that are probably more prevalent today than they were in the past—problems that ought to make one wonder about what sorts of shenanigans lie hidden behind any set of test scores on which one may rely for making educational decisions or erecting theories about the nature of mind. Forty years ago I knew a wonderful sixth-grade teacher (actually a senior colleague whose pedagogical skills I much admired)—a teacher whose ideas about education were so old-fashioned that she honestly believed she could teach her pupils to THINK. She was quite unaware of the high-pitched controversies then swirling through the world of

psychology over the question of nature versus nurture or the constancy versus the inconstancy of the IQ—just as thousands of teachers today, I remind you, are similarly unaware of the same controversies in their latter-day incarnation.

That sixth-grade teacher, however, knew a thing or two about the arithmetic, if not the theory, of intelligence testing. She knew that if you could push up the raw scores on an intelligence test while the pupils' ages remained about the same, then by golly, their IQ's would show a comforting rise. She would have had a ready answer to the vexed question: "Can we boost the IQ?" for she was routinely boosting it annually in her own pupils. Her method was simple. She had got hold of all four forms of the old *Otis Self-Administering Test of Mental Ability*, and, in all good conscience, she used the items of the test as exercises in a unit on intelligent thinking. This she conducted strictly in the drill-and-practice mode of instruction—minus, of course, any aid from a computer, since computers had not yet arrived on the educational scene. The gains in IQ she produced in her pupils were breath-taking in their magnitude and beautiful in their upward flight.

Clearly, that sixth-grade teacher was not playing according to the rules of the testing game. But this was because, like many of her present-day counterparts, she was simply unaware of the rules. On the other hand, had she known what the rules were, she would probably have thought them an irksome constraint on what she regarded as effective teaching. Given her pedagogical frame of reference, she would have had a rather good point. For it was a frame of reference that included the old formula—lately revived in some forms of programmed instruction: "Test, teach, test, teach, test, and teach to mastery." And where else than in those old Otis Tests would she have found such well-worked-out exercises for applying the formula to the teaching of intelligence? The trouble was she got the teaching mixed up with the testing. What she did not realize—and what I am afraid many people still do not realize—is that if you use the test exercises as an instrument for teaching, you destroy the usefulness of the test as an instrument for measuring the effects of teaching.

II

This sampling of events from the remote and recent past is meant to be a reminder of something which is so obvious that, in our preoccu-

pation with the sticky problems of test theory, we sometimes tend to forget it: namely, that testing is first and last a series of *human* transactions and that the problems of testing—even many of the theoretical ones—are essentially *people* problems.

Broadly considered, there are four groups of people who are involved in the transactions we call educational testing: the test *makers*, the test *givers*, the test *takers*, and the test *users*. From this view of the enterprise, two observations can be made. First, both within and across the four groups of participants, there is an extraordinary amount of diversity in their understanding of tests and in their attitudes toward testing. And second, as mass testing has spread throughout the schools of the nation, it has become more and more compartmentalized—that is, disjointed—with the result that the interrelationships among the four groups (the makers, the givers, the takers, the users) have become increasingly strained and tenuous. And the consequence of this is that communications among them are becoming more and more like random events.

It is therefore probably not going too far to say that testing, like so much else in this technologically interdependent society, is characterized by *entropy*, which, according the fourth definition in Webster's, means trending toward "a state of inert uniformity of component elements: absence of form, pattern, hierarchy, or differentiation . . . the general trend of the universe to death and disorder (*Webster's Dictionary, 1965*, p. 759)." This may not be the most cheerful way of putting the case, but, in all seriousness, I suggest that, for the next decade, the overriding problem in the universe of testing is to find ways of reducing the entropy by getting more adequate communication among the human components of testing.

One strategy for beginning to tackle this problem might be to break it up into six pieces by examining the failures of communication that are occurring in each of the six pairs formed by the four component groups. In the rest of this talk, I shall glance at what seems to be happening in just two of the pairs: the one formed by the test *makers* and the test *users;* and the one formed by the test *givers* and the test *takers* in large-scale testing programs.

So how do we get the test makers and the test users on approximately the same wave-length? Among the test makers I include not only the people who write test items and assemble them into whole tests, but also all the backup personnel who plot the procedures for test validation, for scoring the scaling and norming, for estimating

test reliabilities, and so on. Behind them is the small army of test theorists with their various and sometimes contentious notions of how all these matters are to be arranged.

The roster of people who contributed to the Second Edition of the book *Educational Measurement*, which Bob Thorndike has so brilliantly edited (Thorndike, 1971), constitutes a small but fairly representative sample of what I am choosing to call the test-making population. The test *users*, on the other hand, are all those people—teachers, school administrators, guidance counselors, parents, public officials, citizens' commissions, and the like—who use test scores to make decisions of one kind or another, but who, for the most part, *have not read Bob Thorndike's book*.

Indeed, the very excellence of the book attests rather dramatically to the widening gap between test users and test makers. For, as the latter have become increasingly sophisticated in developing the science and art of educational measurement—as indeed they have—the test users, through no fault of their own, are finding themselves ever more deeply in the dark.

The reason for this state of affairs is not far to seek. It lies primarily, I believe, *not* in a deficiency of intellect among the test users, but in the fact that, having become so many, they are now hurting from the effects of Dyer's First Law of Information Dilution, which states that, as knowledge expands while the population of potential users of knowledge also expands, the probability approaches unity that everybody is ignorant of what anyone else knows. In other words, the great majority of test users simply does not have the *time* to look up or catch up or keep up with the enormous number of tests and the mountainous literature on testing that the test makers continue to pile up. (Even some of the test makers themselves seem to be having a bit of trouble in this respect.)

The dimensions of the test-making explosion are suggested by the fact that the ETS Test Collection now holds 680 different tests in just one category alone—reading. And it may be further noted that Oscar Buros had to expand the *Seventh Mental Measurements Yearbook* (Buros, 1972) to *two* fat volumes, supplemented by various interim publications, to accommodate the output of tests and the literature on measurement.

On occasion, in order to try to help people find what I thought might be a short-cut to a reasonably tight definition of their educational objectives, I have made the simple-minded suggestion to citi-

zens concerned about local or state testing programs, that they would do well to undertake a systematic examination of a number of existing achievement tests, *item by item*, to see which tests would best reflect the types of behavior they were hoping for in the children attending their schools. Almost invariably their response to the suggestion has been one of dismay. We are all busy people, they say in effect. How can you possibly be so unrealistic as to suppose that we could find the time to go through such a demanding exercise—even if we knew how?

Their point of course is well-taken. But the problem remains. How does one get the users of tests—especially those who use tests as instruments for determining educational policy—to know enough about the innards of the tests they are using to have some clear ideas of what the test scores are saying about what children are learning and schools are teaching? There has to be an answer to this question, and I think it is incumbent on the test *makers* to find it. One step toward an answer is probably to be found in the efforts of the National Assessment of Educational Progress to get people to focus attention in the first instance not on scores but on the way students respond to specific questions. Another step will be to unravel the mysteries in the criterion referencing of test scores in such a way that the minds of test users will not be forever fixated on norms.

III

Now let's move over to the opposite corner of the testing universe where a somewhat different set of people problems seems to be flourishing. This is the corner where the test *givers* meet the test *takers*, and the ancient problem of test reliability comes back to haunt us in new ways.

During the last 20 years or so there has been a growing concern and a fair amount of research on the test-giving behavior of teachers with diverse cultural backgrounds as it affects the test-taking behavior of pupils with diverse cultural backgrounds—rich, poor, black, white, Anglos, Chicanos, and so on. Although the research on the problem is still spotty, it is nevertheless pretty conclusive in support of the commonsense notion that pluralism in the test givers interacts with pluralism in the test takers in ways that tend to depress the reliability of test data. Which is to reinforce what the theorists have been telling us all along: that the reliability of a test does not inhere solely in the

testing instrument itself, but in the total testing process. That is, it is, to some extent, a function of the multiplicity of human transactions that occur inside the examination rooms where the test givers and the test takers confront one another. The problem, then, is the very practical one of getting a good estimate of the degree to which these human interactions are in fact influencing the reliability of the test scores.

To put the problem in a fairly concrete way, suppose you have a large-scale testing program in reading and math in which 50,000 fourth-grade pupils are being tested by 1,500 teachers in 500 schools serving areas that range from densely urban through richly suburban to sparsely rural. In this situation, how do you monitor the test-giving-test-taking process in such a way as to get some sort of believable estimate of the amount of error variance that will have been contributed to the test results by the 1,500 test givers who come to the task with varying hangups about the children and with varying degrees of defensiveness, reluctance, and suspicion concerning the entire testing enterprise? Similarly, how do you get the necessary information for a believable estimate of how much additional error variance will have been contributed by the 50,000 test takers with their varying perceptions of what the testing is all about, their varying degrees of understanding about what they are supposed to do, their varying motivations, and their varying attitudes toward both the tests and the testers?

Thus, it seems to me that it is within this congeries of human relationships that we have to define the reliability problem as it exists in the real world of testing. I think it is a problem that is badly in need of more attention than it has been getting from the test makers. Failing this, I am afraid that debates over what test data mean and how far they can be trusted in the formulation of educational policy will, like the never-ending debate touched off by the Coleman Report (Coleman et al., 1966), be forever inconclusive. It seems to me ironic that in a 67-page paper on the quality of the data in that Report, Christopher Jencks devotes barely one page to the reliability of the achievement tests that served as the dependent variables, and even so, gives nary a hint of these human aspects of the reliability problem (Jencks, 1972).

There are many other old problems in the universe of testing that may need some reshaping to bring them into line with the human condition. Some of them have been touched on in the papers you heard this morning. The others will have to be your homework for the next

Invitational Conference on Testing Problems. One particular matter that I hope you will look into is that of recycling the test validation problem to fit the case of those 50,000 children in 500 schools who are trying, through a single test channel, to transmit to test users, with diverse informational needs, bundles of messages about their mathematical competence which may differ from school to school because of the differing creeds about mathematics to which different schools subscribe.

In my view, none of the problems I have hinted at is essentially insoluble. Indeed, I'll hazard the guess that the solutions to most of them are lying right now buried somewhere in the technical literature—but not just the psychometric and psychological literature—in which case the immediate task is to get them out of the literature and translate them into terms that will make them functional in school testing programs.

The accomplishment of this task, however, assumes that all parties to the test communication network can acquire sufficient wisdom to break out of the mind set that perceives tests primarily as devices for sorting out test takers to accommodate the rigidities of educational institutions, rather than as instruments for loosening up the institutions to accommodate the developmental needs of the test takers in ways that will enable them to do something to diminish the entropy in this terribly troubled world.

In the luncheon address that Daniel Starch gave at the 1954 Invitational Conference (Starch, 1955), he asked and attempted to answer the question: "How can advances in science be made to produce advances in wisdom?" That question, it seems to me, is as pertinent to the science of testing as to any other field. It reminds me of two searching questions that were put to me recently during a conversation with a Vietnam veteran.

He was a young black who had dropped out of an inner-city high school at grade ten and had enlisted in the Marines. This step, as he put it, had "saved him from the streets." It had also enabled him to pass the high school equivalency test. But it had given him an educational experience—some of it rather grim—beyond anything he might have got had he stayed in school, even though he claimed that he was still a slow reader. His two questions were these:

First, "Do you think that people these days are generally smarter than people were in the old days?"

I fumbled around with that one, and, thinking of the studies com-

paring Army Alpha scores in World War I with AGCT scores in World War II, I said, "Yes, people these days probably *are* smarter than people used to be."

And then he said, "Yeah! But do you think they are any *wiser* than they used to be?"

That's another problem I am leaving for your homework.

REFERENCES

Anastasi, A. (Ed.) *Testing problems in perspective.* Washington, D. C.: American Council on Education, 1966.

Buros, O. K. (Ed.) *The seventh mental measurements yearbook.* Highland Park, N. J.: The Gryphon Press, 1972.

Cohen, D. K. Politics and research: Evaluation of social action programs in education. *Review of Educational Research*, 1970, **40**, 213–238.

Coleman, J. S., et al. *Equality of educational opportunity.* Washington, D. C.: U. S. Government Printing Office, 1966.

Dragositz, A. Testing programs, statewide. In L. C. Deighton (Gen. ed.), *The encyclopedia of education* Vol. 9. New York: Macmillan, 1971. Pp. 182–185.

Jencks, C. S. The validity of the EEOS data. In F. Mosteller & D. F. Moynihan (Eds.), *On equality of educational opportunity.* New York: Random House, 1972. Pp. 446–512.

Kansas State Senate Bill 301, Section 7, 1971.

Kirst, M. W., & Mosher, E. K. Politics of education. *Review of Educational Research*, 1969, **39**, 623–640.

Starch, D. "... And have not wisdom." In *Proceedings of the 1954 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1955. Pp. 96–104.

Thorndike, R. L. (Ed.) *Educational measurement.* (2nd ed.) Washington, D. C.: American Council on Education, 1971.

Webster's new international dictionary of the English language. (3rd ed., unabridged) Springfield, Mass.: G. & C. Merriam, 1965.

Session II

Educational Programs to Accommodate Cultural Differences

Where is Preschool Education Going: or Are We En Route without a Road Map?[1]

IRVING E. SIGEL
State University of New York at Buffalo

Early childhood education (ECE) may well face a new challenge in the 1970s, particularly in its role as an agent for preparing children from disadvantaged backgrounds for competence in school. In the 1950s ECE faced a decline, where nursery schools were phasing out, where students were not enrolling for teacher training; the 1960s saw a resurgence and a headlong pell-mell renascence of ECE, culminating in Head Start-type programs heavily funded and introduced with considerable fanfare. The mission of ECE became remedial: the way to break the cycle of poverty, reduce the chance of educational failure, improve the health and welfare of the newly discovered poor. Early childhood education was cast into the savior role—the institution that would eliminate the presumed intellectual, social, and emotional deficits among children and families from impoverished environments. From 1965, when President Johnson dubbed the Head Start program a solution to a major social problem to date, ECE has been discussed, evaluated, tinkered with, criticized, and in general has been the center of considerable professional and lay interest (Gordon, 1972). Interestingly enough, a large number of psychologists who ten years before might never have thought of identifying with ECE, became deeply involved and committed. The move from the laboratory to the classroom, from the observer to the intervener, from the detached scientist

[1]Special thanks to Dr. Robert Pruzek for his helpful comments.

to the involved practitioner, shows how attractive the movement toward ECE has been (Stanley, 1972). In spite of harnessing some of the best professional talent in the country, in spite of relatively vast sums of money appropriated to the cause and in spite of hours of devoted commitment on the part of parents and of professionals from virtually every relevant discipline, the hopes and expectations of ECE for achieving the defined goals have not yet been fully realized. I am afraid that the 1970s may see a period of disillusion set in, a retrenchment from the exciting, creative period to one of skepticism, doubt, and even retreat from ECE (Horowitz, 1972). Recent surveys and public pronouncements contribute to the swing of the pendulum from great hopes to great disillusion (Jensen, 1969; Jencks, 1972).

The obtained outcomes should not have been too surprising if careful reading of previous research in long-term remedial effects of ECE were done. Swift, in her review of ECE, points out that results were inconclusive for middle-class children; when lower-class children or those from deprived homes were involved, IQ changes were noted (Swift, 1964). This latter group of studies, however, was limited by the uniqueness of the study populations (orphans, abandoned children, for example). The intent of these studies was to show the impact of environment in IQ terms, in an effort to support the optimism of environmentalists, to wit: that environment is the essential determinant in development of intelligence and an appropriate environment can counteract the negative consequences of undesirable environment. The current perspective in ECE has been that early intervention can influence those attitudes, behaviors, concepts, and skills predictive of adequate school functioning.

An increasing number of studies have been showing that these programs have not achieved these goals, and further, cannot work. Early childhood education is in fact not the way to resolve the problem of poor school performance for these children (Jensen, 1969). Some critics have gone so far as to claim that ECE is harmful and that formal schooling should be delayed even longer (Moore, Moon, & Moore, 1972).

The challenge thrown to the advocates of ECE, then, is to justify it as an educational experience in the face of these reports, justify it in terms of cost, of time, or for reorganization of our educational institutions in relation to outcomes. If advocates of ECE fail to offer constructive suggestions now, a premature decline of ECE could result.

The challenge is understandable in a society such as ours where we

justify social programs from an economic perspective. The cost factor often overrides humanitarian goals, and programs however meritorious are rejected as costly in terms of dollars, not in terms of reduction of human suffering or enhancement of human potential. My point is that the value of ECE as a broad-based education for *all* children integrated within the mainstream of the educational mainstream has *yet to be instituted let alone evaluated.* In this paper, I will present arguments justifying ECE as a worthy educational effort.

Early childhood education as a large-scale national effort is a mere decade old and then only geared for children from economically disadvantaged backgrounds. Before rejecting ECE as an educational entity of value, let us review briefly where we were ten years ago, what we had to do, and what we have accomplished in this short decade.

With the exception of the Montessori program, there were virtually no well-articulated, defined, and worked out ECE models for the group served.[2] Interest in the educational aspect was at a low ebb. Relatively little educational research was being carried out with preschool children, and the studies that were done were particularly concerned with personality development and its antecedents. Little attention was paid to learning, cognitive functioning, educability, language acquisition—all areas that have flourished in the past decade. In essence the field of child development in general and ECE in particular was fragmented, limited in scope, atheoretical. Early childhood education was particularly heavily influenced by mental health considerations (Beilen, 1972). In sum, the field was ill-equipped to meet the new challenge—the readying of minority group impoverished children for school.

To meet the new demand to remedy intellectual and social deficits of disadvantaged children, a basic and fundamental reorientation for ECE was needed. New programs had to be built which would be skill relevant; "teaching" had to be in terms of concepts, skills, and behaviors relevant for elementary school.

The only way the field could cope with these new social demands was to bring together its best hunches, the little empirical knowledge available, and much experience. In addition, crash research programs had to be instituted.

---

[2]This is not to say that ECE programs were nonexistent (Bank Street College, The Merrill-Palmer Institute). They tended to be broadly gauged, focusing on socialization rather than on cognitive and language skills. Also they were generally not "packaged" and prepared for the mass education effort demanded by Head Start.

The urgency to establish programs in spite of the lack of knowledge dealing with so-called deficits of the impoverished meant that the task of programming the educational effort was put together with little time for thoughtful planning, coordinating the relevant knowledge from the various disciplines (education, psychology, sociology, anthropology, biology), and pilot testing. It was a crash program where the social demands and services were primary. In effect, programs were really pilot efforts, and hence these evaluation studies should be generally considered as pilot studies. Under these conditions, how could we expect that remediation plans would be clearly articulated and carried out to ameliorate the deficits of varying intensities and enable children to profit from schools in ways *superior* to their peers who did not attend such programs? In addition, the expectation that short-term gains, if any, in these programs could be maintained over extended time periods was a heavy burden, one that required cooperation of the field of elementary education—a type of cooperation not always received. The response to this was disillusionment with early educational intervention as a solution to the "deficit" education problem.

I am not disillusioned by the disillusionment. From the research perspective, the expectations were unrealistic and even naive. We knew our conceptions were inadequate, our measures crude, our research designs fraught with error. In spite of sophisticated techniques and data analytic models, the initial effort should be viewed as a pilot project. There are some exceptions, but even they are full of conceptual and methodological shortcomings (Stanley, 1972).

We ought not to be disheartened about all of these negative research findings for reasons which I shall discuss. From the humanistic point of view, I am not disappointed, since a large number of children and their parents had valuable and interesting experiences, improved nutrition and diet, opportunities for general health examination, and an enrichment of parents' and children's lives. In fact they had experiences which otherwise would have been denied them. In sum, a lot of children and parents had a lot of happy times—a break from the harsh realities of a life of poverty.

Congress and economy-minded groups, however, are not impressed by these seemingly sentimental impressions, so let me turn to perhaps more convincing reasons for my lack of disillusionment.

We have before us a greater body of knowledge about early childhood development than ever before (Mussen, 1970). We have before

us a more comprehensive understanding of educational settings, programs, materials, and teacher strategies (Gordon, 1972). For example, the research of Louise Miller at the University of Louisville has shown how different preschool programs have different effects over varying lengths of time (Miller, 1972). Her research shows that ECE programs are not monolithic in spite of the claim of the proponents. The vast body of material has contributed and can contribute to our understanding of the dynamics of growth. We must not confuse our impatience and disappointment in not creating the expected long-term social change in such a short time with the increasing knowledge base. In fact, we are in a better position to plan now for ECE programs than we were in 1965. In addition to our increase in knowledge, we have increased communications, we have established such communication centers as ERIC, increased participation and communication among various disciplines. Let us not overlook the potential in resource we have harnessed for continued efforts. Thus, we have gained much by this recent impetus in ECE.

I believe that the impact of ECE on the children also is greater than assumed, but empirical proof of this resides in a reevaluation of our expectations and our data, reconceptualization of the concept of impact, and a reinterpretation of evaluation procedures.

First, consider the general expectations assigned to ECE. It was believed that ECE alone could ameliorate severe or serious deprivation—an unrealistic idea. Amelioration of consequences of psychological, social, or nutritional deprivation coupled with poor health care is a major undertaking requiring the concerted effort of many agencies. No ECE program can be so broadly conceived as to undo or redo the the impact of the total environment that is hostile, rejecting, and neglectful. To attain such social goals requires drastic social changes in practices, attitudes, and feelings among the various ethnic and racial groups.

I do not look to ECE as a panacea for social ills. Rather, ECE can provide opportunities for *all* children to engage in educational experiences which at least in the present can extend their knowledge, provide a break from some of the everyday problems at home, and provide a chance to extend their horizons. Such experiences do provide children with opportunities for excitement, a zest for doing, for learning, for enjoying—and *none* of these are particularly undesirable or unworthy objectives.

Further, it is ironic that the expectations we set for elementary

school are not as grandiose as ECE. Do we worry about the long-term effects of kindergarten or grade one? If grade-one experience facilitates grade-one learning, that would be enough. I doubt if we would try to relate grade one to junior high school performance. And if we find no relationship, would we do away with grade one? Why make these types of demands of ECE?

I feel that the reasons we have concluded that ECE programs are virtual failures may in part be due to the way programs were conceptualized and data were analyzed and evaluated. In the remaining section of this paper, I wish to suggest some procedures which I believe will reveal that we have lost much information or impact of ECE because of how we have approached the problem. Let us begin by defining ECE in behavioral terms.

Early childhood education involves engaging children in experiences which can influence the course of their growth. It is essentially a developmental issue. However, I fail to see where educators and programmers conceptualize the entire process in developmental terms. Evidence for this assertion comes from reading current theory and noting evaluation procedures which tend to view development as incremental, cumulative, and linear, overlooking the dynamic integrative nature of development (Werner, 1957). Narrowing of the conceptualization of development to study of change alone is a distortion. It leads to the misconception

> . . . that development is the kind of anarchism where totally unrelated and discontinuous stages follow each other in sequence; or it may lead to the kind of historicism where new or modified behaviors are merely added on in a continuous temporal order as the organism grows [Langer, 1970, p. 733].

Development is a dialectical process in which the organism's orgarization becomes increasingly differentiated and specific. In the course of these processes new integrations occur, so new learnings are increased in the old and the net result is a new organization.

In other words, think of development as a spiral: learning occurs at one level and then is reintegrated into a new level, not just added on. Learning vocabulary is not just adding the word in an associative way but integrating the word in a context.

If development is conceptualized in these qualitative (organizational) terms, it becomes possible to construct measures assessing levels of functioning. For example, we employ a task to determine the

child's level of representation. The task is a four-stage task moving from the concrete to varying levels of representation: graphic, verbal, and so on. Children vary in the level at which they might solve the problem. This type of task can express the concept of development. Thus, the child who succeeds in the four conditions is not incrementally superior, but is qualitatively different on a different level from the child who can handle only the concrete condition. I believe we need more tasks that are constructed in this fashion.

Evaluation of developmental changes, then, requires procedures which tap into qualitative, as well as quantitative, differences in performance. Changes in children's responses to Piagetian conservation of substance tasks are good examples of what I mean. The classical conservation task might well be the prototype. Let me illustrate by a discussion of the conservation of quantity task.

The child is given two equal balls of clay. Once he attests to equality, one of them is altered in shape and the child is asked if they are still equal in quantity. The child does not solve the problem when about four or five. As he grows older, he begins to solve the problem, but he gives different answers, increasingly sophisticated. At first, he says the two balls of clay are the same because one is just flattened out; later he solves the *same* problem but now uses a more sophisticated explanation: nothing was added or subtracted. More complex answers are given later on, such as invoking principles of compensation: as something gets flatter it gets longer, with quantity remaining constant. The change in the child's conception is not evidenced in his answer to the initial question, "Are these two the same?"; rather, it is in the explanation of his answer. Only by this method of interrogation does one discover the qualitative change. This is a prototype of the evaluation I feel is necessary to assess effects of ECE.

Another issue when working within a developmental framework is to examine the interaction of particular achievements. Basically the question is to assess effects of change (or lack of it) in one area on performance in other areas. For example, what is the consequence of enhancement of skills in fantasy or verbal imagery on attention? It is conceivable that high premiums paid for fantasy and imagery might lead children to move away from attending to the specific concrete. Interference between one act of learning on another is important because it may well result in neutralizing potential gains.

Accepting this perspective, investigation will have to examine the interaction of various tests; it will require programmers to concep-

tualize and evaluate interaction effects of various program components. Unfortunately, most programmers either ignore the interaction of intra-organismic subsystems or pay lip service to it. The end result is that little is known about the interactive effects of emphasizing particular program components vis à vis others. For example, what is the effect of language stimulation on social behavior or task orientation on creativity and problem solving?

Thus, program builders and evaluators working within a developmental framework would have to embark on componential analyses within an interactive analysis framework. This moves away from univariate type approaches and also away from heavy reliance on product-oriented tasks. Now, new analytic models have to be created that can handle synergistic relationships and provide data on the significance of particular interrelationships among variables. Treatment effects have to be examined not in terms of change of a single measure —for example, IQ language comprehension—but in terms of pattern changes. For example, are there concomitant changes or are changes in some areas marked by no changes in others? By this type of examination the more subtle and perhaps significant effects of the program can be ascertained.

Interpretation of the results, however, is difficult for four reasons: 1) the type of task employed; 2) the timing of testing; 3) the situational conditions; 4) the mode of analysis.

Tasks frequently are unrelated to program objectives. Rather than achievement-type tasks, IQ or similar measures are frequently used, not necessarily integral to program goals or content; and these measures are not necessarily those which predict to school performance. These measures should serve a "swivel-chair" effect, evaluate consequences of the ECE program, and be predictive of performance in school. The reason is simple. There is no true integration of these two educational experiences. Components for each program have to be identified in variable or dimensional terms and appropriate tasks created. This procedure is not only expensive in time and money, but also speaks for a certain conceptualization. If that acceptance of the significance of this viewpoint is not evident, of course it will not happen in spite of funds or other conditions. This lack of integration, I believe, negates a meaningful relationship between ECE and elementary school. Further, it does preclude an understanding of various developmental processes because development is not necessarily dependent on an artificial classification of children's educational status. Thus, if

a goal of ECE is predictive of elementary school performance, integration is necessary to define the impact of ECE. Construction of achievement tests for program components should be a must.

Evaluation measures are usually cast in a pre-post design. Examination of process—those conditions influencing effects—has been done but not sufficiently to identify relationship between inputs and performance. Of course, such testing calls for careful consideration of ways to cope with practice effects, subject overtesting, and subjects feeling constantly under scrutiny. There is no need to dwell on a point familiar to ETS audiences and those of you engaged in research.

Third, a crucial factor influencing test performance is the situational factor—the context of the evaluation. Cole and Bruner (1972) state the case well by saying

> . . . when we systematically study the situational determinants of performance, we are led to conclude that cultural differences reside more in differences in the situations to which different cultural groups apply their skills than to differences in the skills possessed by the group in question [pp. 175–176].
>
> Psychologists concerned with comparative research and comparisons of social and ethnic group differences in particular must take seriously the study of the way different groups organize the relation between their hands and minds; without assuming the superiority of one system over another, they must take seriously the dictum that man is a cultural animal. When cultures are in competition for resources, as they are today, the psychologist's task is to analyze the source of the culture differences so that those of the minority, less powerful group may quickly acquire the intellectual instruments necessary for success in the dominant culture, should they so choose [pp. 176–177].

Cazden makes a similar point in evaluating language studies (Cazden, 1970). The challenge here is considerable, for we now have to devise ways of interpreting results obtained under these conditions. Now we have to seek new ways of dimensionalizing situations so we can compare performances not *only* in terms of scores, but in how these performances relate to the dimension of the situation.

For example, some situations in which children are evaluated involve time pressure demands, where others are free and untimed. Differential behaviors and competencies can occur as a function of the coerciveness of time. Which is the more accurate statement of ability? Do we not have to qualify the statement in terms of the time dimension?

True, the above considerations are often taken for granted in adult testing. The error committed in that instance does not justify our continuing it with children.

The social context, with its supports, has an impact on performance. How the adult treats the child relative to the child's expressed needs influences performance (Zigler & Butterfield, 1968). We come head on to the issue of standardization and the degree to which this precludes our maximizing information about the child's knowledge.

Fourth, modes of data analysis must avoid premature foreclosure by formal statistical analyses. I believe much is lost by the rush to quantification. At this time we need to inspect consistencies in the data through careful monitoring of them. Relationships between variables can be arrived at by reflective examination of configurations. Identifying patterns of variables for single individuals as well as groups will be conducive to generating more realistically complex statements of hypotheses. An example of this type of hypothesizing is as follows: children who are task oriented, attentive to directions, and have good verbal comprehension will do better on novel tasks involving verbal instructions.

The evaluation of the child must include his experiences if we are to understand the dynamic relationship between ECE program and skills, attitudes, and concepts necessary for later schooling. To do this requires that we reexamine our view of the nursery school itself. If we consider the child a dynamic organic whole, so, too, the nursery school. It is by its very nature an organism—that is, a system, open and fluid. The preschool as an organization is made up of components which interact and intersect. The major components, grossly delineated are: *actors* (teachers, parents, children); *materials* (books, paints, blocks, trips, and so on); *space* (rooms). Each set of actor's roles are predefined. For example, teachers employ strategies with materials in the space to effect certain outcomes with children; the children are the "learners" reacting to or perhaps initiating teacher's behaviors in the same physical setting. The parents' roles are more variable depending on the objectives of the program, but still with definable limits. These roles are rarely overlapping: teachers are not children; children are not teachers; and parents will of course vary but in fact will not be teachers in the same sense as the teachers. Each of these is in a complex organization of definable and partitive wholes.

The object of this type of analysis rests on the assumption that the better we understand the child's place in the system—that what in-

fluences him and what he influences, influences the system—the better able will we be to evaluate impact and significance of program components, individually or in their unity. Acceptance of this agreement requires creation of methods to assess the situation longitudinally and interactionally because we are dealing with a continuously changing organism whose change also influences the system.

When one examines the problem from a developmental perspective, it is necessary to identify significant relationships at any given time and follow this through various transformations, for each segment impinges on subsequent ones. For example, once a child masters color concept, such mastery may influence his performance on puzzles; in effect, each level (or segment) of achievement sets the stage for a subsequent performance. Achievements, however, are not always so contiguous. For example, learning colors *and* improvement in hand-eye coordination tasks may result in more complex block building structures in the present, or later—a week, a month or a year. Thus, change in behaviors as a function of ECE is not necessarily and directly one to one, but may be indirect or long term. To understand, then, the impact of such variables requires the systematic developmental approach advocated here. In a sense, we can learn from the medical model which argues for synergistic effects of phenomenon as well as side effects. Ironically this was an approach employed by those working from a psychoanalytic framework. Perhaps, as we moved toward quantification we moved away from this. With our more sophisticated techniques and computers, it may now become possible for our conception and our data analysis procedure to be more closely aligned.

One of the potential outcomes of this type of analysis is to enlarge our knowledge of human development since it now becomes possible to examine intra- and interorganismic relationships. Granted the complexity of the problem, we have a source of data which would be tapped as a resource in contributing to solution of such critical developmental issues as effects of early experience on later development, pinpointing particulars in this sphere.

I have wandered far and wide to explain why I am an advocate of ECE. Until the data are in, I see no need to be disillusioned. I feel ECE can teach us much about child development; it can provide valuable experiences for children, today—now. It can also have an important influence on education as a whole. Integrating ECE into a school system may also influence subsequent school organization, curriculum and

teaching strategies. Anderson and Shane (1972) say that

> The "domino" effect of ECE on all subsequent schooling must not be underestimated. The new type of clients who enter the primary program probably will create program changes and mutations that will ripple upward through the middle childhood school years into the secondary school [pp. 367–368].

Not only can ECE influence the school situation, it also can have impact on families. It can provide opportunities for children to be cared for in a protected situation while parents have their opportunities for extending their own lives. Granted this is what Day Care is all about; for me, Day Care should incorporate an educational component. When it does, it becomes an ECE program.

So ECE contributes on a number of fronts: scientific, social, and personal.

Before concluding, however, let me leave you with a dream—a pleasant one which would be the marriage of *avant garde* methodologists to substantive developmental theorists yielding broad-scale research models for ECE. I do not believe the past efforts were adequate in time and resources to accomplish the crucial tests. Efforts were aborted resulting in too few, if any, mature offspring. If we are serious about demonstrating ECE as a viable educational system with denotable consequences, we need large-scale, well-supported research.

I hope I have offered suggestions which can be implemented. I have taken the liberty of including an appendix to demonstrate how I have tried in my own program with two-year-olds to practice what I preach.

APPENDIX

Conceptual Analysis of Early Childhood Education Project (*SUNY/Buffalo, New York*)

The discussion will be concretized by a presentation of the Early Childhood Education Project, of which I am director at the SUNY/Buffalo. Briefly, the program was begun in 1969 under the auspices of Office of Economic Opportunity support. Children were enrolled in 1970, at age two and have been in the program ever since. Now to the analysis of the program in perhaps over-simplified system terms.

The theoretical base derives from a Piagetian orientation, with particular concern for the acquisition of representational competence, where representational competence is defined as the ability to construct and reconstruct the environment in terms of symbols and signs (Sigel, 1972; Piaget, 1951). The objective, then, of the program is to create conditions which would foster such development, because representational competence is basic to thought and adaptation to the environment. Some of the areas of interest are reading, translating physical events and objects into signs and symbols, awareness of the congruence and equivalence of pictorial and linguistic representations of identical physical objects. Given these assertions and definitions of the phenomena, the next step was to construct conditions which would foster opportunities for such behaviors to emerge. Still thinking on a theoretical base, I began to try to reconstruct the life experiences that might activate, maintain, and facilitate the representational processes. What I began to do was to extrapolate from the literature, and reflect on that plus my own experiences. These reflections led to the idea that the critical experiences that activate representational skills can be described as distancing experiences, "those which serve to differentiate the environment in time (present-past-future), in space (here-not here), and in appearance as opposed to reality (observable-inferential). Such differentiating experiences have been termed distancing since distance is created between subject and object (Sigel, 1971, p. 26)." The concept of distance as such is not new. Werner and Kaplan (1963) use it in discussing language and symbol formation, and Piaget uses it once. None of the authors use it in the same sense and for the same purpose as proposed here.

Before proceeding to detail the building of the model, I should in passing point out that inherent in the above statement of the problem, a number of assumptions, unstated or unasked, are made, such as: representational competence is a generic competence and is learned; experience activates an organism; there is a legitimate distinction between subject and object. I accept these assumptions without further debate. Let us leave these comments, then, as signals for further consideration and return to the developing program.

The articulation of the above assumptions is more than an academic exercise or pretension. Rather, it is a *must*, because they provide the theoretical base from which to derive the logic of program teaching strategies (activators), materials (appearance-reality), organization of the classroom (time and space). By clearing these issues I am now

ready to undertake the practical problem of training people in the teaching strategies (teacher training), build a curriculum (ordering of events and materials in the program), and place it into temporal-spatial context. These then become, on the gross level, the components of the system. From here the interaction of these components will play out their roles and control can only be exerted within certain limits and under certain conditions.

This model, however, is incomplete and specialized because it has not taken into account other organismic variables; for example, the role of affect, particular linguistic conventions, the quality of support, and reinforcement of articulated representational behavior. Since these variables were not spelled out, the teachers were not certain how to respond to each of these and still be consistent with the system. For example, the teachers had no trouble with the cognitive dimensions of the teaching strategy, but they asked: How does one maintain consistency between cognitive interactions and disciplinary-management issues? Is it possible to discipline or manage a child in such a way as to undo what one is attempting to create in the cognitive sphere? If one teaches that class membership is arbitrary, for example, depending on what attribute is selected as criterial, how does this fit with arbitrary statements of right and wrong made by the teacher in course of interaction with peers? Further, how does one scale the teaching strategies and all the rest of program in accord with the child's capability to assimilate such experiences?

The point that must be underscored is that the translation of theory into practice requires a series of careful steps of derivation. I have gone through the process, finding that the teacher, while understanding the theory and the distancing hypothesis, still needed much guidance in developing appropriate teaching strategies.

To know that we are accomplishing our program objectives requires an evaluation of the child's representational ability. Evaluation is in terms of direct measurement of achievements. If inputs in the program are to foster representational skills, then we measure representational skills. We do not measure IQ, or high jumping, or other more irrelevant matters. The program objectives include skill in imitation (immediate and deferred), imagery, use of referential language as classification, anticipation, transformation of three-dimensional objects into equivalent pictorial and verbal forms. For each of these a task was constructed or borrowed from existing tests. In this way, I am testing the degree to which the child could handle problems in a

new setting with unfamiliar materials, but which involved the same type of process engaged in, in the nursery school. In effect, he was given achievement tests, a standard way in education to assess whether the child is profiting from the curriculum.

Now we come to a most complex task, data analysis. Recall, we gave the child an array of tasks, testing his ability in various components of representational competence.

In addition, data gathered includes observations of child behavior in the classroom and tutorial sessions, rating of the children by the teachers, and observation of the teachers.

With the help of Dr. Robert Pruzek[3] we are developing a system for organizing our data so that they can be visually examined as configurations at the level of the individual child, not just as summary data. This system allows considerable flexibility to construct clusters of data, defined in various ways, for various modes of classification and at multiple time points. Through analysis of various clusters and comparative analyses between individuals we are in a position to establish hypothetical relations for eventual testing the task. This type of approach is motivated by the felt need to incorporate the systemic variables. The procedure facilitates understanding of intra- as well as interindividual relationships in this context of the nursery school. By using logical analysis to form clusters of variables and individuals, we are, in effect, generating structural hypotheses about interactions of individuals and variables.

Finally, I raise one critical issue in all this, and that is the problems related to the sample of children and the design of our experiment. Suffice it to say that we are all plagued with problems of sampling, of sample selection, of sample characteristics, and sample size. Messick and Barrows (1972) tell us not to despair because quasi-experimental design models are possible—advice we followed the best we could. Not only is this a possible solution, but it is only a step in the move toward more replication—which in turn calls for a greater integrated comparative effort.

It is time we planned in this direction instead of crying about what *was* and bemoaning the fact that our educational intervention is not as powerful as we would like.

---

[3]Associate Professor, SUNY/Albany.

REFERENCES

Anderson, R. H., & Shane, H. G. Implications of early childhood education for lifelong learning. In I. J. Gordon (Ed.), *Early childhood education*. The seventy-first yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1972. Pp. 367–390.

Beilen, H. The status and future of preschool compensatory education. In J. Stanley (Ed.), *Preschool programs for the disadvantaged*. Baltimore: Johns Hopkins University Press, 1972. Pp. 165–181.

Cazden, C. The neglected situation in child language research and education. In F. Williams (Ed.), *Language and poverty*. Chicago: Markham Publishing Company, 1970. Pp. 81–101.

Cole, M., & Bruner, J. Preliminaries to a theory of cultural differences. In I. J. Gordon (Ed.), *Early childhood education*. The seventy-first yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1972. Pp. 161–180.

Gordon, I. J. An instructional theory approach to the analysis of selected early childhood programs. In I. J. Gordon (Ed.), *Early childhood education*. The seventy-first yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1972. Pp. 203–228.

Horowitz, F., & Paden, L. Y. The effectiveness of environmental intervention programs. In B. M. Caldwell & H. Ricciuti (Eds.), *Review of child development research*, Vol. 3. Chicago: University of Chicago Press, 1973, in press.

Jencks, C. "Harvard study disputes impact of schooling on future income." (Interview) *New York Times*, September 8, 1972.

Jensen, A. How much can we boost I. Q. and scholastic achievement? *Harvard Educational Review*, 1969, 39, 1–123.

Langer, J. Werner's comparative organismic theory. In P. Mussen (Ed.), *Carmichael's manual of child psychology*. New York: John Wiley, 1970. Pp. 733–772.

Messick, S., & Barrows, T. S. Strategies of research and evaluation in early childhood education. In I. J. Gordon (Ed.), *Early childhood education*. The seventy-first yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1972. Pp. 261–290.

Miller, L. B., Dyer, J. L., Salk, G. C., Bard, E. D., & Kritkansky, R. P. *Four preschool programs: New dimensions and effects*. Progress Report No. 10, June 1971–May 1972. Louisville, Ky.: Psychology Department, University of Louisville.

Moore, R. S., Moon, R. D., & Moore, D. R. The California report: Early schooling for all. *Phi Delta Kappan*, June 1972, 615–621.

Mussen, P. (Ed.) *Carmichael's manual of child psychology*. New York: John Wiley, 1970.

Piaget, J. *Play dreams and imitation in childhood*. New York: W. W. Norton, 1951.

Sigel, I. E. The distancing hypothesis: A causal hypothesis for the acquisition of representational thought. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior, 1968: Effects of early experience*. Coral Gables, Fla.: The University of Miami Press, 1970. Pp. 99–118.

Sigel, I. E. Developmental theory and preschool education: Issues, problems, and implications. In I. J. Gordon (Ed.), *Early childhood education*. The seventy-first yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1972. Pp. 13–32.

Stanley, J. (Ed.) *Preschool programs for the disdavantaged*. Baltimore: Johns Hopkins University Press, 1972.

Swift, J. Effects of early group experiences: The nursery school and the day nursery. In M. & L. Hoffman (Eds.), *Review of child development*. New York: Russell Sage, 1964. Pp. 249–288.

Werner, H. The concept of development from a comparative and organismic point of view. In D. B. Harris (Ed.), *The concept of development*. Minneapolis: University of Minnesota Press, 1957.

Werner, H., & Kaplan, B. *Symbol formation*. New York: John Wiley, 1963.

Westinghouse Learning Corporation. The impact of Head Start evaluation of effects of Head Start experience on children's cognitive development. Preliminary draft of April 1969. Athens, Ohio: Westinghouse Learning Corporation, Ohio University, 1969.

Zigler, E., & Butterfield, E. Motivational aspects of changes in IQ test performance of culturally deprived nursery school children. *Child Development*, 1968, 39, 1–14.

Black Colleges and Black Studies

BENJAMIN F. PAYTON
The Ford Foundation

It can be argued that the central dynamic of American social history is rooted in group conflict and the efforts of the new nation to achieve a workable reconciliation of diverse and heterogeneous group interests. From the very outset, education was not only a critical component of this process of adjustment but was, itself, deeply affected by the pluralism of group needs and group desires.

As much of the history of America from the 17th century up to this very day indicates, heterogeneity among peoples did not and does not imply equality among the contending groups. Taking full advantage of the superior technology of England and Europe, dominant groups in Colonial America saw education as a powerful tool with which to buttress their self-esteem and their social position. Recognizing but fearing the wide pluralism of society in the colonies, education became a prime instrument of cultural change in the direction of greater homogeneity. Catholics and Jews as well as Africans and Indians were viewed with much apprehension in the wilderness colonies of early America.

But it was the use of education as a dynamic instrument of social change with respect to the Indians that created a pattern, the vestiges of which continue to complicate education for American minorities to this very day. While the effort to educate and convert the Indian to European ways of life failed, it left what one observer calls

an inerradicable mark on American life. It had introduced the problem of group relations in a society of divergent cultures, and with it a form of action that gave a new dimension to the social role of education. For the self-conscious, deliberate, aggressive use of education, first seen in an improvised but confident missionary campaign, spread throughout an increasingly heterogeneous society and came to be accepted as a normal form of educational effort [Bailyn, 1960, pp. 38–39].

This tendency to treat the education of minorities as a process of conversion to North European life styles is central to the problems confronted both by black colleges and black studies programs. Reared upon the fundamental assumption that there is nothing positive in the group life of minorities to which an educational strategem can be fruitfully related, most white scholars—from Charles Darwin, to Gunnar Myrdal, to Daniel Moynihan and David Riesman—have seen only pathology in Afro-American history and cuiture. Thus, blinded by the lights of Northern Europe, Gunnar Myrdal (1964) in his monumental work, *An American Dilemma*, writes that:

The whole Southern Negro educational structure is in a pathological state [p. 951].

The challenges faced by black colleges and by the supporters of black studies programs are comprised, in large measure, of this cultural myopia on the part of white Americans and Anglo-conformist black Americans. The challenge to their very existence has been made doubly difficult in recent decades by the fact that racial integration of the schools has been conceptually elaborated within the framework of American constitutional theory which gives no legal recognition to group needs or desires. As Milton Gordon (1964) has observed: "From the legal point of view, there are 190 million discrete American individuals [p. 4]." Therefore, the racial integration of the schools and of the colleges and universities tends to be understood as a process of accepting "qualified" black and other minority individuals into largely white educational structures. Supported by the major philanthropic foundations as well as by the federal government, "talent searches" for outstanding black students and black faculty have been organized and conducted on behalf of the predominantly white universities but, to my knowledge, not a single such talent search has been supported or organized on behalf of black institutions in the interest of the racial integration of the black schools and universities. The whole process of attempting to integrate our schools and universities has been

tainted with the tincture of a subtle form of white liberal racism which "accepts" black individuals but views the black heritage and black institutions as distinctly inferior. This racism of white liberalism has been crassly exploited by both northern and southern conservatives and has led to the humiliation of black educational leadership all across the south in the name of the very process of change that was originally intended to provide a larger measure of dignity to black people in this country.

There was a time—and not so long ago either—when black people and their allies felt that the twin forces of urbanization and the destruction of the legal and quasi-legal pillars of segregation would lead to a meaningful increase of opportunity for all black people. Leaders in the black colleges felt little fear or anxiety about these processes for they naively assumed that efforts to desegregate our society would make full use of the experiences and resources of the black colleges as well as of other institutions in our society. They were therefore unprepared for the assaults upon them that were led by the Riesmans and the Jenckses and the Kenneth Clarks. These observers viewed the black colleges as being purely "the products of white supremacy and segregation (Riesman & Jencks, 1968, p. 469)," and therefore as "historical anachronisms" in the new order of an emerging, integrated society.

Nonetheless a growing number of black and white observers have begun to make a closer analysis of the strategy and objectives of the civil rights movement and of the impact of urbanization upon the patterns of American race relations. One result of more recent social science research on the general question of how people act in groups and, more specifically, in situations of interracial contact, has been to shift away from the atomistic conception of society that tended to ignore the structural and collective dimensions of intergroup behavior. Thus, a growing body of social scientists are giving less attention to "the authoritarian personality," or "the mark of oppression," or even "the nature of prejudice" as meaningful frontiers for social or psychological research. Most research efforts today are based on the theory that society is not an aggregate of discrete individuals but a mosaic of special interest groups differing in prestige and power, led by initiators of action who make critical decisions. In this view, individual personality problems are related more to institutionalized and culturally accepted social patterns than to discrete ego needs.

This shift in focus has brought into view what some call a "new

frontier of race relations." The central reality on this "new frontier" is the behavior of institutions. The concept of "institutional racism" (see especially Knowles & Prewitt, 1969) has achieved prominence among some social scientists and along with it the concomitant problem of how to achieve the social conditions which make possible *group* attainment for American minorities. Within this approach, social policy analysis is directed toward the issues and problems of "deracinating" white institutions by transforming them into agencies of minority group achievement rather than instruments of nonwhite oppression. More importantly, this perspective is able to identify and embrace positive elements in the history and culture of minority groups. The potentialities for power and strength are emphasized more than the elements of pathology and weakness. The salient point is the focus upon institutional behavior and practices with the objective of enhancing *group* achievement rather than looking mainly at individuals. As sociologist Earl Raab (1962) observes:

> the formula may have to be reversed . . . extended individual opportunity may depend finally upon group achievement. This is, hypothetically, the new frontier of race relations [p. 20].

These developments in the realm of social theory have appeared almost concurrently with a growing mass of data which show that the relative position of black people in the economy improved somewhat between 1959 to 1966 while the absolute situation declined. Similarly, Andrew Brimmer, among others, has shown that while the absolute and relative situation of middle-income blacks is getting somewhat better, that of low-income blacks is getting worse. There appears to be developing a growing under-class of black people who are failing to benefit from any of the changes in the American society.

The impact of the increased urbanization of black people in the country is even more depressing. I know of no large-scale public system of education in any one of the major urban centers of our country that has successfully come to grips with the problem of educating large numbers of black and other minority youngsters. Nor is the physical isolation of the urban Negro diminishing. In almost every big Northern city the proportion of Negroes attending all-Negro schools is rising, despite some ingenious and sincere efforts to arrest the process. In housing, more Negroes are moving to the suburbs, but many of them have moved to what are, or will soon become, all-Negro sections.

This quick survey of some of the sociological and economic facts affecting blacks in our country today forces upon us the conclusion that while legally enforced segregation is waning, it is being supplanted by an ethnic-class system in which group solidarities persist in such institutions as the family, the church, social cliques, and in the local community, supported now by custom instead of law, while there is relatively free competition and participation in the economic and political systems of our society.

Confronted with these among other seemingly intractable facts of history and culture, a growing number of black thinkers have reopened with fresh intensity the old debate regarding the purpose and strategy of education for black Americans. Disillusioned with both the conceptual depth and the personal meaning of integration as an intellectual construct and as a social reality, they are giving increasing attention to the internal dynamics of the black community—its problems, strengths, and potentialities. Black higher education is one of the foremost of their concerns, particularly as expressed in the realities and the potentialities of the black colleges and of black studies programs.

Philosophically, these black scholars take the position that the black experience is a fundamentally human experience and, as such, contains the ingredients to support a broad understanding of the nature of things. In its most sophisticated expression, this view has close kinship with the philosophy of Alfred North Whitehead (1948):

> There is no parting from your own shadow. To experience this faith is to know that in being ourselves we are more than ourselves: to know that our experience, dim and fragmentary as it is, yet sounds the utmost depths of reality [p. 20].

Within this philosophical perspective, black colleges and black studies programs are related *only in part* to the necessities of segregation and of urbanization. It is true that black colleges emerged at a time when black people were not welcomed in any of the colleges in the South and in few of the institutions of higher learning in the North. But these colleges are viewed here as more than mere reactions to white racism. A major impulse behind their formation is seen as the desire on the part of black people and their allies to, as W. E. B. DuBois put it, develop

> centers of a new and beautiful effort at human education [quoted by Drake, 1971, p. 847].

Whether this refocusing of vision will aid the struggle for black freedom will depend very heavily on just how faithful it is to the "regimen of fact and logic" in the black community and how sophisticated it is in the use of its awareness that the issues of black education in America are not apolitical but are deeply embedded in historical as well as contemporary group conflicts between white and black Americans.

Very few Americans have a comprehension of the historical significance of black colleges and, therefore, have very little intellectual leverage on the contemporary problems these institutions face, and only the scantiest preparation for beginning to understand the significance of the movement to develop black studies curricula especially in the white universities. It is not that this field of experience has suffered from neglect but, as Professor Bernard Bailyn (1960) observed regarding studies of the early history of American education,

> from the opposite, from an excess of writing along certain lines and an almost undue clarity of direction . . . the facts, or at least a great quantity of them, are there, but they lie inert; they form no significant pattern [p. 4].

Regrettably, the contributions of these institutions to the growth and development of American society are seldom stated. For example, it was these institutions that, almost single-handedly, advanced a totally illiterate people to over 80 percent literacy in the short span of time between 1865 to 1930. More recently, black colleges awarded 80 percent of all of the undergraduate degrees earned by black people in this country in 1968. Over the past four years, the black colleges have graduated an estimated 200,000 students, 80 percent of whom have entered into professional fields of endeavor. A report by the Carnegie Commission on Education (De Costa & Bowles, 1971) sums up this contribution most cogently:

> with few exceptions, whatever the Negroes have achieved in the way of professional entry has been achieved through the Negro Colleges [p. 197].

If one but recalls briefly the tremendous odds against which these institutions had to contend from their earliest inception up to the present day, these and their many other contributions appear to be almost miraculous.

The typologies of Ernst Troeltsch (1960) and Max Weber (1947) have provided us with the concept that institutional types represent

end products of particular social movements. Originating in some form of group conflict, and led initially by charismatic personalities, social movements over time develop the familiar panoply of bureaucratic modes of organization and seek, thereby, to stabilize and make permanent the goals and interests of the original movement. In this view, black colleges represent the end product of the movement for the emancipation of black people and their efforts to achieve full citizenship and human dignity. Similarly, black studies programs can be viewed as an end product of the urban rebellions of the decade of the 1960s and the efforts on the part of students who constituted the "black surge" into the white universities to institutionalize, confirm, and validate their urban backgrounds as a firm part of the university setting (see also Kinnison, 1972).

The future of black colleges and the future of black studies programs will be much brighter, and the education of all Americans, much sounder, if the larger society can begin to look upon these two centers of black educational experience as legitimate social institutions. In part, this is not done presently because not enough Americans understand the role of group power in American life. The myth of the "melting-pot" has obscured for too many of us the critical role of group power in the adjustment of white emigrant groups in this country and has therefore left far too many entirely unprepared to accept the necessity for such strength to develop among black Americans. Black colleges are important symbols of black aspirations and black pride. They function to meet critical psychocultural needs as well as to provide basic educational opportunities for black youngsters. Given the proper amount of support and understanding, they can be powerful allies in the more general search for ways to provide a higher education to large numbers of young people from minority groups.

Similarly, it is difficult to overestimate the importance of the movement for college-based study of the role of black Americans in the social, political, cultural, and economic development in the United States. Black studies departments represent efforts to institutionalize this task. Such programs and departments are fraught with many difficulties and many mistakes have been made. Clearly the efforts on the part of some people in such programs to turn the black experience into a mystique that is so unique it requires a black skin to comprehend it, are simply fundamentally wrong. If the black experience is a human experience, it can be understood and appreciated by persons

from the full range of backgrounds of the family of man. Just as clearly, it will not be accepted as a serious intellectual discipline if it simply becomes the latest accession to the catalogue of urban, ethnic politics. The effort to institutionalize these programs into various kinds of study units and departments represents a realistic assessment of the university as a setting in which it requires organized group pressure to sustain group needs and group interests over any reasonable period of time.

Both black colleges and black studies programs represent efforts on the part of black people to see their history and their culture confirmed and validated with some degree of permanence through the process of institutionalization. The eloquent words of Lerone Bennett (1965) are pertinent to the conclusion of this paper:

> Institutions are great social pools in which men see themselves and their ideals reflected. They are instruments with which men come to grips with the questions: Who am I? Where do I belong? Without meetings, without rituals, without ceremonies, myths and symbols, men cannot define themselves or enter into real relations with others. American Negroes, recognizing this, attempted first to enter institutions formed for Americans—and were rebuffed. They then embarked on a perilous journey of self-naming, self-legitimatization, and self-discovery [p. 43].

These tasks must continue.

Questions and Answers

Q: How do you explain the separatist attitudes which seem to be reenforced among many blacks on predominantly white campuses with established black studies programs resulting from demands of black students? Do you see this as a positive spinoff or an accompanying problem to be reckoned with? If the latter is so, how would you suggest that one cope with it?

A: I think Chuck Willie's new book on black students in white colleges speaks very well to this question. It seems that as the number of black students on the campus increases, so does the tendency to form separatist enclaves. This seems to have something to do with the quality of the institutional environment insofar as it appears to black students not to be receptive of them as whole persons. They perceive the basic environment of the white university as hostile and they with-

draw into the tents of their own group for solace and comfort.

This is a positive spinoff to the extent that it contributes to group self-reliance. Clearly, it is also a defense mechanism. On the negative side, it clearly does not enable the youngster to take advantage of some of the real benefits that would be his or hers were the environment and the black students able to achieve a better relationship.

Q: What long-term pluralistic model do you have to offer to compete with the model of universal fraternity and equality? Please specify.

A: Well, I'm not offering a model to compete with universal equality, universal fraternity. What I am offering is a pluralistic approach to education that makes a genuine equality and fraternity possible. It is too easy to enshrine education in mystical ideals about democratic values while at the same time continuing a program of cultural dominance based upon a narcissistic preoccupation with European cultural values. Can one really love a black student when the ecology of one's emotions and perceptions devalue him culturally? I think not. It's always easy to love man in the abstract, but it's very difficult to love your neighbor. It's easy to love everybody, but it's very hard to love that black kid who's in your class. And I suppose I'm specifying a model that contends with that. I think I'm saying, insofar as institutions go, and the effort to achieve equal opportunity or equity for minorities in higher education, that there's a role for a variety of institutions and a necessity that curricula reflect the pluralism that is America. This *is* a pluralistic society. I think it's time that many of us recognize this, and try to develop pluralistic sets of institutions to deal with it.

REFERENCES

Bailyn, B. *Education in the forming of an American society*. New York: McGraw-Hill, 1960.

Bennett, L. Jr. *Confrontation: Black and white*. Baltimore: Pelican Books, 1965.

DeCosta, F., & Bowles, F. *Between two worlds*. New York: McGraw-Hill, 1971.

Drake, St. C. The black university in the American social order. *Daedalus*, 1971, **100**, 833–897.

Gordon, M. *Assimilation in American life*. New York: Oxford University Press, 1964.

Kinnison, W. A. A model for urban innovation: The land-grant system. *Educational Record*, 1972, **53** (1), 45–50.

Knowles, L., & Prewitt, K. *Institutional racism in America*. Englewood Cliffs, N. J.: Prentice-Hall, 1969.

Myrdal, G. *An American dilemma*, Vol. II. New York: McGraw-Hill, 1964 (paperback).

Raab, E. *American race relations today*. New York: Doubleday, 1962.

Riesman, D., & Jencks, C. *The academic revolution*. New York: Doubleday, 1968.

Troeltsch, E. *The social teachings of the Christian churches*. Two vols. (Trans. by O. Wyon.) Harper Torchback Edition. New York: Harper & Row, 1960.

Weber, M. *The theory of social and economic organization*. (Trans. by T. Parsons.) New York: The Free Press, a division of the Macmillan Company, 1947.

Whitehead, A. N. *Science and the modern world*. New York: Mentor Books, 1948.