21.
Differential Psychology: Towards Consensus

ARTHUR R.JENSEN

For a researcher to see a specially composed collection of critiques comprehending virtually all the main themes of his own contributions and influence, by a number of the world's luminaries in the relevant fields, is a rare privilege indeed. I am indebted to Drs Sohan and Celia Modgil for initiating this project, and I am especially grateful to the outstanding scientists and scholars who have contributed their expert views of particular aspects of my work. I have been delighted to find much more than mere summaries of my publications in these chapters, and also much more than narrow technical critiques resembling the reports typically expected of referees of journal articles. Although the essays here are clearly focused on the main themes of my work, they are nonetheless highly original and creative contributions in their own right. The intentionally critical bite of some of these essays makes this volume far different from the typical festschrift, which it most definitely is not, and which I would much less prefer.

I have read all the chapters at least twice, the first time through as I might casually read just any book, and the second time more slowly, with a more thoughtful and analytical attitude. The second reading was much more rewarding; I found many riches I had not grasped the first time. This was especially true of the more broad-gauged and philosophic chapters, such as those by Bereiter, Brand, Gordon, and Havender, which expressed some profound insights I have not thought about as much as I have thought about the topics in those chapters that deal with more familiar viewpoints and criticisms or with the prevailing specialized and technical issues that occupy me almost every day in my research. I can happily say that, all told, I have found this volume of essays more richly rewarding than I had expected. Perhaps over the last seventeen years, since my work first became 'controversial', I have become jaded by the plethora of often superficial and ill-informed commentary. It is most refreshing to find something very different here, in technical competence, breadth, thoughtfulness, and cleverness. In these essays can be found none of the Laurel and Hardy quality of some of the earlier commentaries by unqualified critics, none of the patently ideological polemics, and, with what seems to me only one exception, virtually none of the *ad hominem* brickbats that only vitiate what is supposedly scholarly criticism.

Although this is not the proper place for autobiography (see Jensen, 1974), I think it appropriate, in commenting on each of the main topics in this book, to say something about how I got into the topic and why it seemed interesting and important to me. But before getting down to specifics, I should mention a few rather general tendencies that may have colored my behavior as a researcher.

I have never been much of a believer. I have felt little need for belief *per se*. Even as a child in Sunday school, I aroused complaints for my too persistent doubting, questioning, and arguing with the Sunday school teacher. Agnosticism in the most general sense was always more natural and congenial to me than passive acceptance of what others believed. I always liked to question, to seek evidence, to look for consistencies. Hence the concept of truth, emphatically spelled with a small 't', that is, truth in the scientific sense, when I first learned of it, had great appeal to me. The appeal has not waned. Early on I seemed motivated to question popular myths and entrenched beliefs. Whatever emotional needs I had for any kind of subjective certainty I found entirely in unarguable aesthetic experiences, in my strong involvement with music. A sense of morality that requires no supernatural justification was conditioned by my rather strict parents and later instilled at some higher level of consciousness, since about age 12, by my hero worship of Mahatma Gandhi. I first read about him in *Time* magazine and then proceeded to read everything available about him in the school library and in the public library. I even went so far as to become a vegetarian for a time, to my parents' consternation, and to write a book-length biography of Gandhi and also to edit a book-length compilation of selections from Gandhi's writings that I considered most representative of his thoughts on a variety of subjects. (These rather immature efforts were never published.) My fascination with the great life and character of Gandhi has continued to this day, although, of course, I have long since had to restrict severely my reading in this area, as the Gandhi literature now is vast—ninety volumes of his collected writings and more than 400 books about his life and work.

When, as a high school student, I first read a book on psychology—J.B. Watson's *Psychology from the Standpoint of a Behaviorist*, given to me by an aunt who had used it in one of her college courses—I was attracted not only by Watson's lively style and the interesting subject matter itself, but by Watson's iconoclastic stance, so unlike the various bland elementary science books I had been reading haphazardly since I was in grammar school. Thus began my interest in psychology, which I later pursued as a student in college. Until then, however, it was greatly overshadowed by my interest in music.

Although always a reluctant believer, I have not been at all averse to inventing and pushing hypotheses, which I regard only as a means of testing reality and finding new facts. Even a seemingly zany hypothesis, if rightly worked, may serve a useful purpose. I feel no kinship with those intellectually inhibited skeptics whose chronic reaction to almost any novel hypothesis is quick dismissal. My one and only indispensable postulate, I suppose, is the existence of an objective reality (with a small 'r') underlying observable phenomena which, in

principle, can be understood in the scientific sense through human ingenuity. It seems to me that without this fundamental postulate, scientific pursuits would be futile—at best a game, which could hardly compete with other intellectually, artistically, or socially productive pursuits for one's major lifelong commitment.

Also, I am probably prone to a certain naïveté in my approach to phenomena, even at times being most amazed by 'the obvious'. But this tendency may be of some advantage to a researcher. Armed with few preconceptions, one encounters many phenomena that are amazing and puzzling and invite investigation. But practical limitations also force one to be selective.

The several personal proclivities that have influenced the selection of the kinds of problems on which I have done research have been described more fully elsewhere (Jensen, 1982a). Briefly, I am attracted by phenomena which have already accrued popular explanations, or unquestioned beliefs, or have spawned contradictory theories. A phenomenon's interest is also enhanced if it seems counter-intuitive, surprising, or inexplicable in terms of any established principles. Those phenomena that more readily lend themselves to reliable measurement or have potentially quantifiable properties are also more attractive subjects for scientific study. A phenomenon with fairly robust and regular or 'lawful' aspects, as contrasted with one requiring extremely specific conditions for its manifestation, is also a more likely prospect for fruitful investigation. A psychological phenomenon that more directly seems to have biological underpinnings or more clearly suggests it could be a product of human evolution in the biological sense is generally more interesting to me than the predominantly cultural aspects of behavior. Finally, I am attracted by unresolved problems that are deeply rooted in the history of psychology. The nature of human mental abilities and the measurement of individual and group differences in intelligence are topics that quite completely meet this perhaps idiosyncratic combination of proclivities. Many phenomena in this domain evince all the features of attractiveness for investigation that I have indicated. Besides these attractions that are intrinsic to research, there is the added bonus that the subject of human intelligence is commonly viewed as having crucial relevance to education, to society, and to human welfare. In recent years an increasing number of psychological researchers, many among the contributors to this book, have come to recognize this subject's rich potential for scientifically rewarding and socially significant research. I hope that my own activity in this domain has enhanced its visibility and contributed to the increasing recognition of its importance.

HUMAN LEARNING AND THE LEVEL I/LEVEL II THEORY

My research activities can be divided conveniently into pre- and post-1969, the year that the storm of controversy arose over my article 'How Much Can We Boost IQ and Scholastic Achievement?' in the *Harvard Educational Review* (Jensen, 1969). The present book deals almost exclusively with the post-1969

phase, which follows my public introduction into what has been popularly termed 'the IQ controversy'. But this now famous article was my seventy-sixth publication, as I had already been publishing in psychological journals over fourteen years. Julian Stanley, in his Introduction, refers to a 'little-known initial article (Jensen, 1968a)', published a year before the *Harvard Educational Review* article; this 'little-known' 1968 article was probably my first fully intentional attempt to point out the significance of certain major topics in differential psychology for education. At the time it did not seem to me a 'little-known' article, as Stanley has characterized it, since it drew over 700 reprint requests (a record for me at that time) and was reprinted in two books of readings. But then it was almost completely forgotten after the appearance of the much more highly publicized 1969 *Harvard Educational Review* article, which carried essentially the same message, although in a much more elaborated way.

Until 1969, however, nearly all my work was in the field of human learning—learning theory, particularly of the Hullian variety, and the classical problems of serial and paired-associate verbal rote learning. (What I did in that field is succinctly summarized elsewhere [Jensen, 1974].) The Level I/Level II conception grew out of this work in 1960, when I began testing economically disadvantaged Mexican-American children, many in classes for the educably mentally retarded, on tasks consisting of serial and paired-associate rote learning and free recall of familiar objects, however they were labeled by the child (Jensen, 1961). I was struck by the considerable disparity between the level of performance on these 'direct learning' tasks, as I then called them, and scores on conventional IQ tests. Most Mexican-American children in retarded classes did as well on the learning tasks as their Mexican-American and Anglo-American age-mates in regular classes. On the other hand, most of the Anglo-Americans in retarded classes performed at a much lower level on the direct learning tasks than did their age-mates in regular classes. This interaction of ethnicity, IQ, and rote-learning ability seemed important because it suggested that Mexican-American and Anglo-American children who were identified as the educably retarded on the basis of standard IQ tests, and were put into the same special classes and treated alike, were actually quite different in their learning capability and therefore should probably receive quite different educational treatment. There was no real question that, scholastically, all were failing in regular classes. But was it for quite different reasons? If so, quite different treatments might be indicated.

I tried out these direct learning tasks in other groups: low, average, and high IQ; low and middle socio-economic status (SES); black and white children. Both race (black-white) and SES showed the same kind of interaction with IQ and rote-learning ability that I had found with Mexican-Americans. Subsequent studies, however, showed a much less pronounced interaction for different levels of SES within ethnic groups than for black and white groups, even when they were matched on standard indices of SES. This phenomenon, I concluded, pertained more clearly to differences between typical black and white children than to differences between social classes *per se* or to the only other ethnic groups on

which we had data: Mexican-Americans and Asian-Americans (Chinese and Japanese). Asians, in fact, showed a slightly opposite kind of interaction to that seen in the black-white comparisons, that is, Asian children performed slightly higher on IQ, particularly non-verbal reasoning, and slightly lower on rote learning and short-term memory, as compared with white children of the same age. It was especially clear in our studies that representative samples of black children were much less different from their white and Asian age-peers on tests of serial and paired-associate rote learning and short-term memory ability than on IQ tests, whether verbal or non-verbal. The most striking examples were found among black children in retarded classes, with IQs authentically below 75. It was not uncommon to find that their forward digit span memory and serial rote-learning ability were in the average (or above) range for white children in regular classes. This great disparity between IQ and digit span was seldom found for white children in retarded classes.

The explanation of such findings, I hypothesized, was the existence of two fairly distinct classes of abilities, which I called Level I and Level II. I have never thought of Level I/Level II as a 'theory', although it has been called that. Actually, it is scarcely more than a simple empirical generalization describing the interaction of three types of variables: (1) a limited class of memory and learning abilities, (2) IQ or performance on similar complex cognitive tasks, and (3) certain population dichotomies. To call this a theory seems too grandiose; if I use the term 'theory', it should be understood in quotes. *Level I* ability involves the accurate registration and recall of information without the need for elaboration, transformation, or other mental manipulation. It is most easily measured by *forward* digit span memory and serial rote learning of verbal material with minimal meaningful organization. (*Backward* digit span memory requires transformation [reversing the order of the digits in recall], and therefore shows a higher correlation with IQ and a larger mean black-white difference.) *Level II* ability involves transformation or manipulation of the input in order to arrive at the correct output—reasoning, problem-solving, inference, semantic generalization, conceptual categorization, and the like. Level II is virtually the same as Spearman's construct of *g*, the general factor common to all complex tests of cognitive abilities. Level II is probably even closer to what Cattell refers to as 'fluid' *g*, as contrasted with 'crystallized' *g*. However, I have not regarded Levels I and II as factors of ability in the strict factor analytic sense, but as two categories of tasks requiring either a fairly minimal or a fairly large amount of transformation of information for successful performance. The relatively low intercorrelation between Level I and Level II tests and the extreme cases in which there is high Level I despite very low, even retarded, Level II, indicates distinct abilities underlying these categories of task performance.

Because we more often found high Level I ability in the presence of low Level II ability than the reverse combination, I had hypothesized a causally hierarchical relationship between Levels I and II, such that the development of Level II ability is dependent on Level I. That is, adequate Level I ability was

seen as a necessary but not sufficient condition for the development of normal Level II ability. There is statistically significant evidence for this hierarchical relationship, but the relationship appears weak and cannot be viewed as an important aspect of the Level I/Level II formulation.

Level I/Level II was important, I think, because it revealed a type of mental ability in which black-white differences are minimal as compared with the ability (or abilities) measured by traditional IQ tests or similar highly *g*-loaded cognitive tasks. It suggested what then seemed a promising possibility, that the Level I/ Level II distinction might lend itself to an aptitude-by-training interaction that could decrease the disparity in scholastic performance between typical black and white children. This hope has not panned out, I conjecture, because of the intrinsically highly g-loaded, or Level II, nature of educational achievement. Educational achievements seem to be valued almost directly to the extent that they are perceived as *g*-loaded, and this is as true for any ethnic minorities in our schools as for the white majority.

In recent years I have placed less emphasis on the Levels hypothesis, which I now view as merely a special case of what I regard as a much broader and more fundamental phenomenon that Spearman first noted in 1927 and which I have termed 'Spearman's hypothesis'. This hypothesis states that the black-white difference is essentially a difference in g, and the varying magnitudes of the mean black-white differences (in standard-score units) on various tests are directly related to the tests' *g* loadings. A preponderance of evidence substantiates Spearman's hypothesis, although there is also evidence that certain other factors independent of *g*, such as spatial visualization ability, also show mean black-white differences, but to a much lesser degree than the *g* difference (Jensen, 1985a, 1985b; Jensen and Reynolds, 1982). Hence Level II can be equated with Spearman's *g* and Level I represents only a fairly narrow category of tasks (rote learning and memory span) among all those tasks that show especially low loadings on *g*. I still think it worthwhile to investigate the broad realm of very low *g*-loaded cognitive tasks in relation to various population differences, with a view to discovering abilities that may afford some educational and occupational leverage for *individuals* who fall markedly below the norm of performance on highly *g*-loaded tasks. Because of what now amounts to a virtual moratorium in the United States on research on racial differences in psychological traits, this line of research seems unlikely to receive any concerted effort in the foreseeable future. The prevailing attitude is to deny the reality of *g* or of population differences in *g*, or to hold out hope of markedly raising the level of *g* in disadvantaged groups by some purely educational means, as yet undiscovered.

Vernon. This essay is a good, straightforward review and defense of the Level I/ Level II formulation and in conclusion points out the Levels theory's transitional nature in the development of a more complete understanding of population differences in abilities. I myself would emphasize my more recent view of the Levels theory as merely a special case of Spearman's hypothesis, as I

have previously indicated. Also, Vernon's paper does not reflect my now somewhat lesser optimism about the possibilities in Level I for the educational and occupational advancement of those who are disadvantaged in Level II ability, or *g*. I have not yet found evidence that other traits, independent of *g*, can actually *substitute* for *g*, when *g* is below some minimal threshold required for successful performance, a threshold that varies, in a probabilistic fashion, for different levels of education and occupation. Provided that an individual's *g*, or general ability, exceeds this prerequisite threshold, other traits—special talents, motivation, persistence, dependability, character, etc.—may very significantly enhance the individual's chances for success. But *g* has some significant degree of predictive validity for quality and efficiency of performance in virtually every type of job in our society above the level of the most unskilled labor. Moreover, I find no real evidence that different populations possess different kinds of intelligence that are substitutable on a par for *g* as we know it. I doubt that the black population's leaders are asking for recognition cf a different *kind* of intelligence, with different consequences for educability and employability in this society. What they want is the same distribution of success rates as in the white and Asian populations, in school, in college, and in the job market—success rates that are importantly related to the psychologist's construct of *g*, and this relationship is the same for blacks as for whites and Asians. That seems to be the real problem. I have no illusions, as it is equally clear that Vernon does not, that Level I offers anything that could ameliorate this condition.

Stankov. This essay seems bent on refuting the Levels theory by any possible means. With a few exceptions, most of the arguments and the data they are based on are weak or faulty criticisms of the Levels theory. Stankov claims 'there is a large body of evidence contrary to the theory', but he does not marshal this evidence and at times even misconceives the theory to score a point. Vernon (1981) has reviewed virtually all of the relevant research on the Levels theory prior to 1981, which is the bulk of the evidence, and in his present chapter he updates this review. In no way does there emerge from *all* the relevant studies anything in the least resembling the claim that 'there is a large body of evidence contrary to the theory', to quote Stankov. The preponderance of the evidence is quite consistent and quite the contrary of Stankov's claim.

The purpose of Level I/Level II was never 'to function as a theory of the organization of the whole broad range of human cognitive abilities', as Stankov suggests, but merely to describe one salient aspect of the nature of SES and black-white differences in cognitive performance. More recently this formulation was limited almost exclusively to black-white differences.

Stankov's arguments seem to presuppose that factor analysis is the only way to study ability differences. If groups differ systematically in the specificities of tasks (i.e., the reliable test-score variance not included in the common factor variance), this would not be revealed by factor analysis. But it would be revealed by classifications of various tests in terms of their discrimination between certain population groups that are of particular interest for educational or occupational

reasons. Those instances in which group differences show up in certain empirically derived *categories* of tests whose factor loadings do not line up the same on the major dimensions of ability as revealed by factor analysis would be of special theoretical and practical interest. The Levels theory was never at all in *competition* with factor theories, such as those of Guilford or Cattell and Horn, that attempt to embrace the entire abilities domain in terms of factor models.

Contrary to Stankov's assertion, I do not recall ever having classified or ever having thought of figure-drawing and figure-copying as tests of Level I ability. It has long been clear to me from our studies using the Gesell Figure-Copying Test that a child's ability to copy a given geometric figure is dependent on being able to *conceptualize* the figure in terms of its most essential features, which is obviously a Level II ability. It is the child's abstract conceptualization of the figure, rather than the sheer visual image of it, that directs the child's attempt to copy it. The act of copying is essentially *constructive* rather than merely *reproductive*. Within the narrow psychometric constraints (for example, range restriction and mediocre reliability) of the typical figure-copying tests, they are quite good measures of *g*, with *g* loadings comparable to Raven's Matrices in some of our studies. Also, figure-copying is one of the tests on which we have found the largest ethnic group differences in young children, consistent with their differences on other Level II or highly *g*-loaded tests. On the Gesell Figure-Copying Test, for example, a large representative sample of *fourth*-grade black children was found to perform, on average, on a par with Asian and white *first*-graders (Jensen, 1973b, pp. 304–5).

Stankov is unfairly reluctant to acknowledge that all tests that have been classified as Level I are not equally good measures of Level I; forward digit span and serial rote learning of nonsense syllables, for example, are more pure Level I than backward digit span or memory span for meaningful words or sentences, which may permit greater conceptual organization that aids recall.

Stankov shows a profound misunderstanding of my conception of Level I if he thinks that I could possibly classify Vocabulary and Information tests (along with Memory Span) as Level I tests! I have never thought, written, or suggested anything of the kind! As I have clearly explained elsewhere (Jensen, 1980a, pp. 145–7), acquisition of vocabulary is mainly a process of inferential inductive reasoning and not a process of rote memory. Vocabulary is largely acquired by encountering words in a context from which their meanings can be inferred—clearly a Level II function. Information is learned, retained, and retrieved largely in terms of understanding within a framework of organized knowledge and concepts, again Level II. Vocabulary and information acquired by rote learning have little *g* loading. This has been the sad fate of attempts to raise IQ by 'teaching to the test'. The artificially raised IQ no longer reflects *g*, the most active ingredient in the IQ's predictive validity for scholastic and job performance.

Stankov is mistaken again. The Jensen and Reynolds (1982) article does not at all represent the Arithmetic subtest of the Wechsler Intelligence Scale for

Children— Revised (WISC-R) as a Level I test. Arithmetic is shown to be loaded about twice as much on the *g* factor (Level II) as on the short-term memory factor (Level I). Problem arithmetic *per se* is highly Level II, but a short-term memory factor has moderate loadings (0.25–0.32) in the Arithmetic subtest because the arithmetic problems are given orally in the WISC-R, with the possibility of only one repetition, and the testee must therefore be able to recall the essential elements of the problem in order to solve it. As a consequence of its moderate loading on the memory factor, the Arithmetic subtest shows a mean black-white difference only two-thirds as large as Block Design, which has the same *g* loading for blacks as Arithmetic, but a negligible loading on the memory factor. Thus, contrary to the impression created by Stankov's reference to this study, the results are, in fact, highly consistent with the Levels theory. The fact that Level I, or short-term learning and memory, is not just the *absence* of *g,* but is a reliably measurable ability in its own right, is indicated by the finding that when *g* (or the Full Scale IQ) is partialled out of all the WISC-R subtests, blacks perform *better* than whites on just those subtests with salient loadings on the memory factor: Digit Span, Arithmetic, Coding, and Tapping Span (Jensen and Reynolds, 1982).

I have never said that memory span tests *always* have the lowest loadings on the general factor, and, in fact, a number of my own studies show other Level I tests with even lower *g* loadings, for example, Tapping Span (or Knox cubes) and serial rote learning, but these tests are not as practicable for general use or large-scale studies involving group testing.

Stankov's assertion that 'almost all SES differences on G_c [crystallized intelligence] can be explained as due to *g*' is not inconsistent with the Levels theory and is perfectly in agreement with my views on the nature of the correlation between *g* and SES in modern Western societies, namely, that *g* is a predominant causal factor in social mobility via educational and occupational attainments.

Stankov refers to Boyce's study (done in 1983 but still not published) claiming that analyses of WISC-R data at the *item level* in black and white samples frequently appear to contradict the Levels theory, the reasoning type of *items* often showing smaller differences than items involving memory and prior learning (does this mean Vocabulary and Information?). Item-based studies, however, are quite unsuitable for this kind of investigation, because no *single item* has sufficient reliability adequately to reflect anything that would be called a general trait or factor. The *g* loadings of single items are very small compared to their specificities. It should not be forgotten that, even within item-homogeneous tests, the average correlation between items is of the order of +0. 10 to +0.20. The average inter-item correlation in the Raven Matrices, for example, is only +0.12 in the general population. When items of a particular *type* are grouped into a subtest, their largest common factor can be measured with fair reliability. When black-white differences on various such subtests are compared, it is consistently found that the differences are larger on the Level II, or

reasoning-type tests, than on the Level I, or rote-memory-type tests. In the entire WISC-R national standardization sample, for example, the mean black-white differences (in *z*-score units) for Comprehension, Block Design, and Object Assembly are 0.94, 0.93, and 0.82 respectively, as compared with Digit Span, Tapping Span, and Digit Symbol (Coding), with black-white differences of 0.31, 0.33, and 0.47 respectively. It seems to me that even a single study based on the massive WISC-R national standardization data is due more weight than a 'box score' of the results of a number of much smaller studies, many with questionable samples, such as those reviewed by Boyce and on which Stankov relies.

Stankov's mistakes are by now almost tiresome, as well as astonishing. He claims that my 1980 study (with Inouye) is supportive of the Levels theory 'only after corrections for unreliability were applied to the raw scores.' But no 'correction' of any kind was applied to either the raw scores or the factor scores, and both types of scores yielded highly similar results fully consistent with the Levels theory! Then Stankov mentioned a study by Osborne as if it, too, contradicts the Levels theory, but I have never cited any study by Osborne in support of the Levels theory! (There is, however, an explicit developmental study of the Levels theory by Jensen and Osborne [1979], and it is mainly consistent with the Levels theory.)

Stankov's Table 3 and the argument based on it are misleading. The third-order General factor properly corresponds to Level II. But this is a Schmid-Leiman factor analysis, and so the second-order factors, G_c and G_f, are residualized, that is, the general factor, G, has been partialled out of G_c and G_f. These residualized second-order factors are hence diminished in their Level II properties, so the *residualized* G_f (with *G* removed) cannot be accepted as a good Level II marker. Yet we are left with the misleading impression that the black-white difference on G_f factor scores is relatively small. Also, the marker test for the Primary Factor, Auditory Immediate Memory, it should be noted, is a test of *backward* digit span, which involves transformation of the input and usually has about double the *g* loading of forward digit span—hardly an ideal measure of Level I. Consequently, it shows a larger black-white difference than should be expected on a true Level I factor. By contrast, the lower part of Stankov's Table 3 shows the results of Jensen's (1973c) study, in which G_c and G_f were extracted as first-order factors and were not residualized (as they are in Stankov's analysis), and the Memory factor is based on three somewhat different tests of *forward* digit span—a proper measure of Level I. These results are just what one would predict from the Levels theory.

Stankov: 'If group differences exist mostly on the general factor…, one should not abandon other factors in favor of it.' Have I abandoned other factors?

If the Levels theory is just a special case of Spearman's hypothesis, as I now claim, then Stankov's conjecture that the Level I/Level II interaction with race (and with SES) may exist with children but is absent in adults is certainly wrong. Spearman's hypothesis is every bit as clearly substantiated in samples of black

and white adults as in child samples (Jensen, 1985a). Stankov's critique indeed proves to be far less sturdy than the Levels theory and Spearman's hypothesis.

GENETICS OF HUMAN ABILITY

My route into the genetics of human mental ability has been described previously in the Preface of my book, *Genetics and Education* (1973a), and in a brief professional autobiography (Jensen, 1974). The gist is that my interest in this subject was almost inadvertent. It came about in 1966 in the course of my preparing to write a book on the school learning problems of children described as 'culturally disadvantaged', the term in popular usage at that time. The prevailing attitude in the field then was either to ignore genetics completely or to deny its relevance to the study of individual and group differences in mental abilities or other traits germane to scholastic performance. Anyone who doubts this has simply forgotten the history of that period—the zenith of what has later been termed 'naïve environmentalism'. Although, as an undergraduate, I had taken a course in genetics, the subject was never brought to bear on behavior, and the term 'behavioral genetics' had not yet been conceived. It was possible in the 1940s and 1950s, to earn BA, MA, and PhD degrees in psychology without ever coming across such concepts as genotype, phenotype, and heritability. American psychology was almost completely dominated by the behaviorist-environmentalist philosophy. The fact that this viewpoint had become virtually a dogma in the 1950s and 1960s, especially in the areas of clinical, educational, experimental, and child psychology, goes a long way toward explaining the incredible commotion among psychologists, social scientists, and educationists that immediately followed the appearance of my 1969 article in the *Harvard Educational Review,* probably the first major publication in over twenty years to emphasize explicitly the relevance of genetics for understanding certain increasingly prominent problems of public education.

I had begun by trying, for the sake of scholarly thoroughness, merely to write a short chapter for my book on the 'culturally disadvantaged' that I expected would succinctly review the so-called nature-nurture issue only to easily dismiss it as being of little or no importance for the subsequent study of the causes of scholastic failure and success. I delved into practically all the available literature on the genetics of intelligence, beginning with the works of the most prominent investigator in this field, Sir Cyril Burt, whom I had previously heard give a brilliant lecture entitled The Inheritance of Mental Ability' at University College, London in 1957. The more I read in this field, the less convinced I became of the prevailing belief in the all-importance of environment and learning as the mechanisms of individual and group differences in general ability and scholastic aptitude. I felt even somewhat resentful of my prior education, that I could have gone as far as I had—already a fairly well-recognized professor of educational psychology—and yet could have remained so unaware of the crucial importance of genetic factors for the study of individual differences. It was little

consolation that I had been 'in good company' in my ignorance of genetics; in fact, that aspect of the situation seemed even more alarming to me. I was overwhelmed by the realization of the almost Herculean job that would be needed to get the majority of psychologists and educators fully to recognize the importance of genetics for the understanding of variation in psychological traits. Hence, rather than attempting at first to add small increments of original empirical research to the body of knowledge on the genetics of human abilities, I thought my most useful role at that point was a primarily didactic one. Most of my thirty-five articles and four books dealing with genetics are of that nature. But in the course of marshaling the scattered existing research evidence, and trying to make the most sense of it, I noted certain methodological problems and formulations that called for criticism and reformulation. One was Karl Holzinger's conceptually muddled index of heritability based on monozygotic (MZ) and dizygotic (DZ) twins, for which I substituted a more defensible formula that comes closer to the theoretical definition of heritability and also takes account of assortative mating in estimating the heritability of a trait (Jensen, 1967). Another was the estimation of the limits of genotype×environment covariance in IQ, based on data from MZ and DZ twins (Jensen, 1976b). A theoretical paper on the possible explanation of race differences and a race×sex interaction in spatial ability in terms of sex-linkage of a hypothesized recessive gene that enhances spatial visualization ability (Jensen, 1975a), although an interesting and plausible theory, has been undercut in recent years by the failure to find consistent evidence for any sex-linkage in the genetic conditioning of spatial ability. My empirical findings in behavior genetics have concerned the heritability of memory span (Jensen and Marisi, 1979) and the effects of inbreeding depression on general ability (Agrawal, Sinha and Jensen, 1984; Jensen, 1983b). The study of inbreeding depression seems to me especially important in the study of human abilities, because inbreeding depression indicates genetic dominance, and the presence and degree of dominance are related to natural selection for the trait in the course of its biological evolution. It was of great interest to me to discover, for example, that of the several ability factors that can be extracted from the various subtests of the Wechsler Intelligence Scale for Children, the one that shows the greatest susceptibility to inbreeding depression is the *g* factor (Jensen, 1983b). This finding indicates that one of our most widely used standard psychometric tests of intelligence yields scores that reflect some part of the variance in the biological intelligence that has developed in the course of human evolution.

This is a good place to straighten out a common misconception about my position on the heritability of intelligence, a misconception shared by neither Bouchard nor Plomin, nor probably by any other of my readers who are knowledgeable about behavior genetics. I refer to the naive belief in the indelible rubber-stamp identification of my name with an IQ heritability of 0.80, *specifically* 0.80, as if this particular figure were intended as an inexorable constant, from which any bona fide deviation found in other research intimates

an error in my figure of 0.80. This figure apparently comes from my 1969 article in the *Harvard Educational Review*. I had assembled the median correlations of virtually every type of kinship correlation reported in the literature up to that time —a total of fifteen medians of different kinship correlations. With the help of a professor of quantitative genetics, I extracted an overall estimate of 'broad heritability' (h_B^2) from the fifteen median kinship correlations. (The h_B^2 is a statistical estimate of the proportion of the total variance in some measurable phenotype—in this case IQ—that is attributable to all the genetic factors that have conditioned the phenotype.) By means of a 'biometric genetic model' (closely akin to the analysis of variance), the best estimate of the value of h_B^2 was 0.77, and, assuming an overall test reliability of 0.95, the value of 0.77 was corrected for attenuation (i.e., errors of measurement), yielding a corrected h_B^2 of approximately 0.80. (A correction for attenuation is entirely proper for theoretical purposes, since the question of theoretical interest is the proportion of the true-score variance in measured intelligence that is attributable to genetic variation.) I clearly pointed out that the specific value of the heritability would vary due to sampling error, the nature of the population sampled, and the particular tests used to measure intelligence. I later learned that three famous geneticists (among them one of the great pioneers of quantitative genetics, Sewall Wright), each applying somewhat different methods to estimating broad heritability from these same data, arrived at values within ±.02 of my (uncorrected) value of 0.77. Hence the value I claimed was not a figure I just happened to pluck out of thin air, as some critics would make it seem. The estimation of heritability is like any other kind of statistical estimation. It depends on empirically justifiable models of gene action and also on appropriate empirical data. From this viewpoint I have never claimed anything about the heritability of intelligence that was not warranted. To describe me as a 'hereditarian' is ridiculous if this label implies anything other than the fact that I advocate application of the best available methods of quantitative genetic analysis to the study of individual differences and that I accept the findings based on a preponderance of the resulting evidence as scientific fact. Any critic who talks as if I have ever insisted on tenaciously maintaining a special brief for some particular value of the heritability of intelligence is so utterly naïve as to disqualify himself at the outset. Readers who are abreast of recent advances in human behavior genetics will not be at all surprised that nothing at all naïve or in the least misinformed can be found in the expertly instructive commentaries by Plomin and Bouchard.

Plomin. This is an excellent review of the main issues and my approach to them, bringing up to date the most important developments in the genetics of mental ability. Within the past decade or so behavior genetics has burgeoned into a substantial discipline, with its own journal, *Behavior Genetics* (for which both Plomin and Bouchard serve on the editorial board), its own Behavior Genetics Association, and graduate programs in this speciality. There is increasing recognition that behavior genetics is an essential discipline for research in differential and developmental psychology and is indeed germane to all of

psychology and sociology. What I have termed the 'sociologist's fallacy' is essentially the failure to recognize that the observed correlations between environmental factors and human behavior are often largely mediated by genetic factors. This is seen, for example, in the significantly differing correlations between home environmental factors and children's IQs for children reared by their biological parents and for children reared by adoptive parents. The effect of the environment on development cannot itself be properly studied without also taking account of genetic sources of variance. This is largely what human behavior genetics is all about. In my view a psychology without roots in biology and not availed of the methods of behavior genetics is, scientifically, a hopeless enterprise.

Bouchard. At the beginning of his essay Bouchard views my own work as a 'direct extension' of what he terms the 'British biological-theoretical tradition of research in individual differences', which originated with the work of Sir Francis Galton (1822–1911). I think this is an accurate and, to me, complimentary perception. It applies not only to my interest in the genetics of mental ability, but to my more general interest in human psychological variation, its immediate causes as well as its evolutionary history, and its meaning for modern society. Also, my interests in the psychological characteristics of both the extremes of the 'normal distribution', and my theoretical and methodological approach to the study of individual differences in intelligence, using chronometric analysis of elementary cognitive tasks, seem distinctly Galtonian.

But it was really not until some time after I first discovered my natural affinity to Galton's s thinking that I began to study his works seriously. I had previously known of Galton only second-hand, through my study of the history of psychology. It was not until still later, just a few years ago, that I came to discover, quite by accident and with surprise, another kind of 'kinship' with Galton, even more direct than I would have guessed. But, after all, psychology does not have a very long history. A popular science writer had asked me in an interview if I had ever traced down my own 'PhD genealogy'. That is, who was my major professor while studying for my PhD, who was his, and so on, back to the very beginning of the PhD degree in psychology. At the time I did not know my 'genealogy' for more than two 'generations' back, but I later looked into it. Going from the present to the past, the 'direct line' of my 'academic ancestry' goes back four 'generations': Percival M.Symonds, Edward L. Thorndike, James McKeen Cattell, and Wilhelm Wundt, who founded the first psychological laboratory, which, at the time Cattell was a graduate student, was the only institution in the world granting a PhD in psychology. There is also an important collateral branch on this 'family tree'. After receiving his PhD under Wundt, J.McK. Cattell spent a postdoctoral year in London working with Galton, whose intellectual influence on Cattell was notably greater than Wundt's. (Cattell wrote much later in his life that he regarded Galton as the greatest man he had ever met.) Another collateral branch traces back to Galton from my own postdoctoral fellowship—two years in H.J.Eysenck's Psychology Department in the Institute

of Psychiatry of the University of London. Eysenck had earned his PhD in psychology at University College, London, under Sir Cyril Burt, who had studied psychology under William McDougall at Oxford. Burt's father was Galton's physician and, as a youth, Cyril Burt came under the personal influence of Galton. (In the last two years of Burt's long life I became quite well acquainted with him personally, visiting him several times for lengthy discussions each summer I spent in London in 1970 and 1971 [see Jensen, 1983a].)

These connections all seemed uncannily surprising to me when I first noticed them, because I had long before found this particular group of psychologists among the most interesting and congenial I had come across in all my reading. Even as a student of learning theory, years before I became involved in the 'IQ controversy', I was especially attracted to E.L.Thorndike and read his major works on the experimental psychology of learning with great pleasure and, at times, excitement. I cannot be sure how much these historic figures have affected my own career in psychology, but I suspect that the noted affinities were more a result of some predisposition on my part, rather than a directly causal influence.

Bouchard's essay is not only a trenchant critique of the critics of the hereditarian research program, but a constructive general methodological criticism of the hereditarian program itself. Bouchard makes a creative contribution in sketching ways that the genetic analysis of human abilities can be strengthened scientifically in its future course. Behavior genetics is not a fossilized methodology whose future consists of no more than the applications of a static methodology to a catalog of psychological traits of interest. It is a rapidly evolving complex methodology for coordinating genetical models with experimental and psychometric approaches to the understanding of variation in behavioral characteristics. The introduction of rigorous and objective statistical methods of meta-analysis, as Bouchard points out, should make it possible to consolidate the results of many behavior genetic analyses so as to yield conclusions with a degree of resolution and certainty that are unattainable by any single study. As Bouchard states, 'Models allow us to treat the [various kinship] data as a whole, rather than in arbitrary bits and pieces. Meta-analysis helps us understand the data in detail, but prevents us from becoming overwhelmed by artifacts.' It would be hard to find a more succinct statement of a working philosophy for the future course of behavior genetic studies of human abilities. That Bouchard's prescription of 'models and meta-analysis' as 'a set of tools and a set of attitudes that Galton would have been the first to apply in his own laboratory' is suggested by Galton's own words, which appear on the frontispiece of every issue of the *Annals of Human Genetics,* founded by Galton in 1909:

> General impressions are never to be trusted. Unfortunately when they are of long standing they become fixed rules of life and assume a prescriptive right not to be questioned. Consequently those who are not accustomed to original inquiry entertain a hatred and horror of statistics. They cannot

> endure the idea of submitting sacred impressions to cold-blooded verification. But it is the triumph of scientific men to rise superior to such superstitions, to desire tests by which the value of beliefs may be ascertained, and to feel sufficiently masters of themselves to discard contemptuously whatever may be found untrue.

TEST BIAS

Almost every one of the many topics discussed in my article, 'How Much Can We Boost IQ and Scholastic Achievement?' (Jensen, 1969), has been pursued in my subsequent research, to fill gaps in the exisiting knowledge, to answer criticisms, and generally to strengthen the scientific basis for understanding the issues raised by this article. Virtually everything I have done as a researcher since 1969 has grown out of these core issues concerning the nature and measurement of intelligence, the characteristics and causes of individual and group differences, the natural development and changeability of intelligence, its relation to physical and other non-cognitive variables, and its educational, economic, and social correlates.

The subject of cultural bias in mental tests was mentioned only too briefly in the 1969 article, in a short paragraph and with a rather oblique reference to the first article (Jensen, 1968b) I had written on this topic. Soon thereafter, I realized that my article had slighted what was likely to become a key issue in the study of racial differences on mental tests. As research discredited various popular theories of the mean black-white IQ difference—unequal educational opportunity, teacher expectancy, level of aspiration, nutrition—the blame would increasingly be directed at the tests themselves, or at the conditions of testing, such as the effects of the race of the tester and of prior practice on similar tests. These arguments needed more thorough examination than could possibly have been afforded them in the 1969 article.

My first opportunity to expand on the issue of culture bias was an article solicited by the *Toledo Law Review* (Jensen, 1970). At that point I began collecting all the research material I could find on this subject. My interest in the subject dates back to 1950, when I became a student of Kenneth Eells, who was one of the major pioneers in the study of culture bias in standardized tests and under whom I did my master's thesis in psychology. But it was not until more than twenty years later that I felt the necessity for doing research on test bias myself. I began by comparing the rank order of item difficulty (percentage failing) in various standardized tests, such as the Lorge-Thorndike IQ test, the Peabody Picture Vocabulary, the Wonderlic Personnel Test, and the Raven Matrices, in large samples of blacks and whites. The extremely high correlations between the rank order of item difficulty across the two racial groups were highly inconsistent with the culture bias hypothesis, unless one made the most unlikely assumption that all the highly diverse items were *equally* biased. My activity in this area finally led to my writing *Bias in Mental Testing* (1980a).

It began with my intention of writing a small book, briefly explaining the main psychometric issues and methods then in use for investigating bias and summarizing the results of my own studies, several of which had already been given detailed presentation in a number of journal articles. As I went more deeply into the subject, however, I saw that test bias could not be properly understood separately from broader issues in psychometrics and the theory and measurement of intelligence. So the typescript of the book gradually expanded, almost of its own accord, to become fifteen chapters totalling about 1300 pages, which shrunk to about 800 printed pages in the published book. I doubt that I could ever have written even the first page of it if I had been warned at the outset how many pages I would eventually end up having to write.

I regarded this task as something to be put behind me, so as to get on with more basic research in differential psychology. The study of test bias has now become one of the highly technical specialties in the field of psychometrics, and although I keep up with most of the newer developments by perusing the relevant journals and books, since I am still often expected to express expert opinions in this field, the pursuit of further technical refinements for detecting ever smaller and subtler kinds of bias that may have statistical but virtually no practical significance does not greatly interest me. Clean-up operations are still no doubt needed, as well as the routinization of bias detection by all institutions that construct and publish standardized tests. But it seems to me unlikely that, from here on, there will be any radical innovations in test bias methodology or any very startling findings or conclusions that will contradict the evidence now in hand, much of which is included in my *Bias in Mental Testing.* It is noteworthy that some two years after the publication of *Bias,* a panel of nineteen experts, commissioned by the National Academy of Sciences and the National Research Council, reviewed much the same body of evidence on test bias and arrived at essentially the same main conclusions that I had arrived at (Wigdor and Garner, 1982). Hence my investigation led to what is now a well-established and generally accepted position among the experts in psychometrics. My own latest thoughts on the subject are presented in the final chapter of *Perspectives on Bias in Mental Testing,* edited by Reynolds and Brown (1984). In it I state:

> More than 100 reviews, critiques, and commentaries have been addressed to my *Bias in Mental Testing* since its publication in January 1980. (A good sampling of 27 critiques, including my replies to them, is to be found in the 'Open Peer Commentary' in *The Behavioral and Brain Sciences*, 1980, *3*. 325–371). It is of considerable interest that not a single one has challenged the book's main conclusions, as summarized in the preceding section. This seemed to me remarkable, considering that these conclusions go directly counter to the prevailing popular notions about test bias. We had all been brought up with the conviction that mental ability tests of nearly every type are culturally biased against all racial and ethnic minorities and the poor and are slanted in favor of the white middle class. The

> contradiction of this belief by massive empirical evidence pertinent to a variety of criteria for directly testing the cultural bias hypothesis has revealed a degree of consensus about the main conclusions that seems unusual in the social sciences: The observed differences in score distributions on the most widely used standardized tests between native-born, English-speaking racial groups in the United States are not the result of artifacts or shortcomings of the tests themselves; they represent real differences—*phenotypic* differences, certainly—between groups in the abilities, aptitudes, or achievements measured by the tests. I have not found any critic who, after reading *Bias in Mental Testing*, has seriously questioned this conclusion, in the sense of presenting any contrary evidence or of faulting the essential methodology for detecting test bias. This is not to suggest that there has been a dearth of criticism, but criticisms have been directed only at a number of side issues, unessential to the cultural bias hypothesis, and to technical issues in factor analysis and statistics that are not critical to the main argument. But no large and complex work is unassailable in this respect. (Jensen, 1984a, pp. 531–2)

Gordon. This is an absolutely masterful contribution. I doubt that there is another sociologist, or many psychometricians for that matter, with as comprehensive and profound a grasp of the fundamental issues concerning test bias as it relates to racial and ethnic group differences as that displayed by Robert Gordon. Also, I believe that my efforts in this area have never been more well understood or more brilliantly explicated by any other commentator on this topic. Indeed, Gordon's essay is itself a major contribution to the literature on test bias.

The key theme in Gordon's chapter, that lends it theoretical coherence, is his clear perception that the guiding force in my own work in mental measurement arises principally from my constant search for *construct validity* that can embrace the widest range of phenomena in differential psychology. In my philosophy, science is an unrelenting battle against ad hoc explanation. No other field in psychology with which I have been acquainted has been so infested by ad hoc theories as the attempts to explain social class, racial, and ethnic group differences on various tests of mental ability. My pursuit of what I have called the Spearman hypothesis (Jensen, 1985a), which is nicely explicated by Gordon, represents an effort to displace various ad hoc views of the black-white differences on psychometric tests by pointing out the relationship of the differences to the *g* loadings of tests, thereby bringing the black-white difference into the whole nomothetic network of the *g* construct. It is within this framework, I believe, that the black-white difference in psychometric tests and all their correlates, will ultimately have to be understood. Understanding the black-white difference is part and parcel of understanding the nature of *g* itself. My thoughts about researching the nature of *g* have been expounded in a recent book chapter (Jensen, 1986b). Enough said. Gordon's chapter speaks for itself,

and, with his three commentaries on the chapters by Osterlind, Shepard, and Scheuneman, leaves little else for me to add to this topic.

Osterlind. This essay focuses on my treatment of what I have termed *internal* indices of test bias, which includes item bias, a subject on which Osterlind has written an informative book. His expertise in this field and the dispassionate objectivity of his approach to the subject makes me take his few criticisms and points of disagreement with me quite seriously. I find myself essentially in agreement with Osterlind on these critical points. He is one critic, among several others, I wish had been able to review my *Bias in Mental Testing* before it went to press. It would have made for some revisions and improvements.

I must now agree that the Guttman scale represents an unrealistic ideal for mental test construction and is actually unnecessary if there is a sufficient number of all positively intercorrelated items. The *g* factor, which is the core of general intelligence, is best measured by tests composed of fairly heterogeneous items involving a variety of informational content and cognitive processes. But such item heterogeneity precludes a Guttman scale. It would indeed be a major technical *tour de force*, assuming it would even be at all possible, to construct a test of items that conform to a Guttman scale and one that is also as *g*-loaded as such highly heterogeneous tests as the Binet and the Wechsler, when it is factor analyzed among a large and diverse battery of mental tests. Actually, the predominant amount of specificity in single items of any type militates against their being unidimensional; the non-specific components of any important breadth, such as *g* or the major group factors, are so overridden by specificity at the item level as to make the achievement of a Guttman scale practically impossible.

I also agree with Osterlind that item response theory (IRT) deserves a larger place in a book on test bias than I gave it. The basic idea of the IRT technique for assessing bias was given in about three pages of *Bias* (pp. 442–5), with references to the major reviews of the then extant literature. My main reason for not going into IRT further at the time I was writing *Bias* was that, to be useful, an exposition that went beyond the three pages I allotted would have necessitated a full chapter of highly technical material. A lengthy theoretical exposition hardly seemed warranted in view of the scant empirical applications of IRT at the time. The few reports of its application that I could find were mainly intended as methodological demonstrations, without any very systematic results or conclusions of a genuinely substantive nature. Since then, IRT has become the major and preferred method for detecting item bias when very large subject samples are available to the investigator. Certainly, an updated revision of my book would have to include a much more thorough exposition of IRT and its empirical results in recent studies of item bias.

Shepard. Gordon's comments on Shepard's chapter largely obviate the main points I would make. Shepard seems to strain at casting doubt on the validity of my conclusions about test bias. She does this partly by trying to make it appear that I allow no exceptions, no open questions, no loose ends. It is almost like

taking exception to the statement that π=3.14, because π is really 3.14159265, etc. *ad infinitum*.

Anyone reading all of *Bias* will clearly see that I have not treated the absence of bias as an 'absolute' (to use Shepard's term), and that the brief summary statements quoted by Shepard from the very last page of the book cannot reflect all the technical subtleties and qualifications found in the detailed consideration of evidence in the book, although my overall summary accurately reflects the conclusion most reasonable persons would draw from the overwhelming preponderance of evidence, especially as it relates to the currently most widely used tests of aptitude and achievement. In a number of places I have discussed the distinction between indicators of bias, such as item×group interaction, that may be statistically significant, thereby *rejecting* the null hypothesis (i.e., no bias), and yet may be so trivially small and inconsistent in direction across various items, as to have no practical significance, or to come anywhere near to accounting for the overall mean difference in test scores between populations. This was my point in proposing what I termed the Group Difference/Interaction ratio, or GD/I (*Bias*, pp. 561–5), which expresses the magnitude of the variance between groups in relation to the variance associated with the items×groups interaction (the indicator of item bias), and I showed that in the case of the black-white difference the GD/I ratio is very large for such diverse tests as the Wechsler, Peabody Picture Vocabulary, and Raven.

The two chapters of *Bias* (Chs 10 and 11) in which I review the evidence on external and internal indicators of bias make the following summary statements. I ask, do they seem as inexorably 'absolute' as Shepard's characterization would lead one to believe?

> It seems safe to conclude that most standard ability and aptitude tests in current use in education, in the armed forces, and in employment selection are not biased for blacks or whites with respect to criterion validity and that the little bias that has been found in some studies has been in a direction that actually favors the selection of blacks when the selection procedure is color blind. (p. 515)
>
> All the main findings of this examination of internal and construct validity criteria of culture bias either fail to support, or else diametrically contradict, the expectations that follow from the hypothesis that most current standard tests of mental ability are culturally biased for American-born blacks. (p. 587)

Probably the broadest summary statement in the whole book is the second paragraph of the Preface:

> My exhaustive review of the empirical research bearing on this issue leads me to the conclusion that the currently most widely used standardized tests of mental ability—IQ, scholastic aptitude, and achievement tests—are, by

> and large, *not* biased against any of the native-born English-speaking minority groups on which the amount of research evidence is sufficient for an objective determination of bias, if the tests were in fact biased. (p. 14)

Are these general summarizing conclusions peculiar to me alone? Compare them with the more recent conclusions of three leading researchers on test bias and personnel selection, Hunter, Schmidt and Rauschenberger (1984):

> This chapter focuses primarily on the employment domain, where ability tests are used to predict job performance. The massive data now available from test validation studies in that domain show clearly and unequivocally that tests have no bias in measuring ability. In particular, a minority person with a low ability-test score will, on the average, perform just as poorly on the job as would a majority worker with the same low score.
>
> Because tests are not biased in the employment domain, they cannot be biased in any other domain. Evidence suggesting bias would have to have some other explanation. However, we know of no domain where there has been cumulative evidence suggesting bias. (p. 42)

> The hypothesis that cognitive tests are unfair to minority test-takers has been repeatedly subjected to empirical test in studies of job performance. Massive empirical evidence has now accumulated showing that tests are fair to minority members; the mean job performance of minority and majority members is the same when people are matched on the composite-ability test score that best predicts performance. For hiring purposes, this means that minority applicants with low ability-test scores later have the same low job performance as majority applicants with the same ability scores. Massive evidence in the educational domain not reviewed in this paper shows the same thing; minority students with low ability scores do just as poorly in learning situations as do majority students with the same low scores. (pp. 93–4)

When Hunter and Schmidt and their co-workers were criticized for their strong conclusions, much as Shepard has criticized me, their reply seems eminently applicable in the present context:

> This objection questions our style of stating research findings and conclusions. It does not question the actual findings and conclusions. We feel that when there is a large amount of empirical evidence supporting a conclusion, and there is little or no empirical evidence to the contrary (for example, as with test fairness or validity generalization), the 'strong' statements of conclusion which this objection correctly states that we make are not only appropriate but in fact are scientifically mandated. They are not 'flashy' or 'one-sided.' Under these circumstances, weak conclusionary

> statements do not accurately reflect the known facts. Without going into detail, we note here that, if one looks back at the history of science, one finds that scientists have not traditionally spoken or written in a hedged-about, overqualified way. There is little support in the history of science for an overly qualified style of communication. Instead, what we are dealing with here is an unfortunate and aberrant tendency that has developed in I/O [Industrial/Organizational] psychology and some other social science areas over the last few decades. It is not a 'characteristic of science.' (Schmidt *et al.*, in press)

Shepard criticizes my rejection of the concepts of 'capacity' and 'potential'. But I have found no operational or empirically determinable definitions or measurements of these vague concepts as they relate to individual differences in mental ability. The closest substitution I know of that is also a well-recognized scientific construct with operational meaning is *genotype*. This construct is an absolutely essential feature of any scientifically adequate account of individual differences. Given this, the terms 'capacity' and 'potential' have no defensible scientific status that I know of in differential psychology, and they can be easily dispensed with. Either these concepts should be formulated in a scientifically adequate manner, or they should be discarded. Their loss does no damage whatever to the idea that aptitude and intelligence tests are measures of 'developed abilities', which is to say measures of *phenotypes*. We can operationally speak of the correlation between phenotypes and genotypes. (The correlation is simply the square root of the heritability.)

Shepard states that within the test validity paradigm, 'legitimization of the criterion variable represents a value choice.' True. But is the value choice at odds with the values of the groups for which the determination of predictive bias (or its absence) is at issue? Is the criterion not legitimized by the aspirations of the individuals and groups who want to succeed in school, in college, and in the job market? The question of bias or unfairness scarcely arises for groups that do not seek the rewards of the criteria that tests are intended to predict. Blacks and Hispanics, as a group, do not object to the *criteria* predicted by tests, such as scholastic performance, college grade-point average, job proficiency, and the like; they object to the fact that they perform, on average, less well on the *tests* than other groups. Tests are the issue, not the criterion. The troublesome answer found through massive research—that the tests predict the criterion as well for minority groups as for the majority—is well summarized in the previous first two quotations by Hunter *et al.*

Finally, I must note two serious misrepresentations in Shepard's chapter. First, Shepard states that 'Jensen said not only that the inferiority of blacks was real, but that it was permanent, fixed in the genetic code.' This, of course, is a flagrant travesty of anything I have ever said on this topic. Here is what I actually wrote in my 1969 *HER* article:

> So all we are left with are various lines of evidence, no one of which is definitive alone, but which, viewed all together, make it a not unreasonable hypothesis that genetic factors are strongly implicated in the average Negro-white intelligence difference. The preponderance of the evidence is, in my opinion, less consistent with a strictly environmental hypothesis than with a genetic hypothesis, which, of course, does not exclude the influence of environment or its interaction with genetic factors. (p. 82)

I have never used the word 'inferiority' in this context but have always referred to statistical differences in performance on specific variables. The differences are 'real' in the sense that they are not attributable to test bias. Nothing has ever been said about differences being 'permanent', but I have pointed out the failure of educational interventions markedly or durably to raise the *g* component of mental ability. The idea that race differences in *g* are 'fixed in the genetic code', to quote Shepard, is, so far as I know, a hypothesis that no scientist has ever even suggested. (The genetic code is the sequence of nucleotides composing the molecular structure of the DNA that constitutes a single gene.) It would not be true even for such indisputably differing features as skin color! Polygenic theory formulates racial differences in continuous traits in terms of differences in the frequencies of certain genes, not in terms of differences in the genetic code itself. In fact, the modern scientific definition of race is based on the criterion of population differences in gene frequencies, and populations' gene pools differ in only a small fraction of all their genes. But this small fraction accounts for the many racial variations we observe. (Even more than 90 per cent of the genes in the human species are identical to the genes in the anthropoid apes.) A polygenic hypothesis posits that the very same genes that produce variation among persons of the same race can produce variation between races or other population groups that are relatively segregated reproductively. The gene pools of such groups may possess different frequencies of the genes that affect a particular trait. Thus, whatever racial genetic differences may exist in important human traits are statistical rather than typological.

Second, to support her claim that tests have been 'instruments of racism', Shepard states that 'Goddard (1913) administered English IQ tests to foreign-speaking immigrants arriving at Ellis Island, New York, and concluded that the majority were feebleminded.' The article (Goddard, 1913) cited by Shepard clearly does not substantiate her accusation. This shabby slander against Goddard in recent years has become a popular canard by antagonists of psychological tests. What Goddard (1913) actually said was that of those immigrants screened at Ellis Island who were suspected by medical examiners and others of being 'feebleminded' on the basis of casual observations, a majority scored in the 'feebleminded' range on certain verbal and performance tests, including the Binet, which were given in the subject's native language through an interpreter. The 'majority' of those tested who had subnormal scores were among only those who were previously suspected of mental deficiency. They were in no way a

representative sample of the many immigrants going through Ellis Island, the vast majority of whom never were given mental tests. Nor was a random sample of any national group of immigrants ever tested. The only study by Goddard involving the testing of immigrants begins with the following sentence: This is a study not of immigrants in general but of six small highly selected groups, four of "average normals" and two of apparent "defectives," all of them steerage passengers arriving at Ellis Island' (Goddard, 1917, p. 243). (The trumped-up charge that Goddard was 'racist' has been well countered in an article in the *American Psychologist* by Franz Samelson [1982], a historian of psychology.)

Scheuneman. Again, Gordon's commentary anticipates the main criticism I would make of this essay. The basis of it is Scheuneman's apparent reliance on what I have termed the 'sociologist's fallacy', that is, the attribution of environmental causation without controlling for the causal effects of genetic factors. It begins in her first paragraph: '…it would be surprising indeed if obvious differences between racial and ethnic groups in economic advantage and opportunity for learning and advancement had no impact on the development of mental abilities.' This is strictly no more than a statement of correlation, but its wording ('impact on') produces the impression of a causal relationship. Careful readers will find other instances of the sociologist's fallacy in Scheuneman's paper, as well as acceptance of research by sociologists Mercer and Blau which, Gordon notes, are classic flagrant examples of this fallacy.

Early in her paper Scheuneman introduces a straw man into the discussion, claiming that my argument takes the form: 'Score differences occur. Hence if the groups are different in the abilities being measured, the test must be unbiased.' I wish Scheuneman had substantiated this claim with some direct references to my work. I cannot find any basis for it. I believe I have pointed out that if there is strong evidence independent of a test that two groups differ in a trait purportedly measured by the test and the test does not reflect this difference, it may be suspected of bias. That is, bias need not always be in the direction of exaggerating differences; it may also have the opposite effect.

Scheuneman claims I have ignored the findings from item bias research. I did not ignore the findings that were available at the time I wrote *Bias;* nearly all the recent studies mentioned by Scheuneman were published since then. Indeed, a valuable feature of Scheuneman's chapter is its review of many test bias studies done since the publication of my book. But the important question is, do these more recent studies substantially contradict the conclusions drawn from the evidence I reviewed prior to 1980? No one has made a compelling case that they do.

Perhaps even more at the basis of Scheuneman's critique than the sociologist's fallacy is that she has not sufficiently heeded what I emphasized as the necessity for making a clear distinction between the test bias question and the 'nature-nurture' question. My insistence on this distinction is not just window dressing to shield test bias research from the opprobrium leveled against any but strictly sociological hypotheses of racial differences. It is theoretically and

methodologically crucial that bias in measurement not be confused with other sources of variance in test scores. Knowledge of the *causes* of group differences is not a necessary precondition for reaching valid conclusions about the degree of bias in the *measurement* of differences. Because raw measurements of any observable human characteristic are strictly phenotypic, the whole issue of test bias can be, and should be, dealt with independently of questions of environmental and genetic causation. The question of test bias, however, is a crucial aspect of the causal question. Before one even begins to think about causes of differences, it must be established that the phenotypic differences are not merely an artifact of biased measurement. The study of bias is the attempt to answer either one or both of two main questions: (1) Is a test's predictive or criterion validity the same for groups *A* and *B* for whatever use is made of the test? and (2) Do the various means for demonstrating a test's construct validity yield essentially equivalent results for groups *A* and *B*?

The argument that the rank correlation (or other forms of correlation) between item difficulties in two groups is an insensitive index of item biases seems to have no basis other than the fact that, in empirical studies and studies in which biased items are artificially created to test this index, as Scheuneman has done, the correlations fall within a quite narrow range toward the high end of the scale. They are not dispersed over the entire range of possible correlations, i.e., −1 to +1. But does that fact prove they are an insensitive index? Is the clinical thermometer an insensitive index of body temperature because its scale does not range all the way from absolute zero to the temperature of a blast furnace? The absolute size of the correlations is not as important as whether the relative magnitudes of the correlations accurately reflect differences between tests in their amounts of item bias. The item correlation, however, is just an overall indicator of item bias; it does not pinpoint the specifically most biased items, for which other methods, such as comparison of item characteristic curves, are appropriate.

I doubt that Scheuneman's notion of a 'constant degree of bias' across items has any possible operational means of detection. It amounts to an untestable ad hoc hypothesis to account for group differences when no evidence of test bias can be detected. Because the idea of a 'constant bias' across all diverse test items can be conceived in the abstract is no evidence for its existence. It is a unicorn.

One can probably always detect some statistically significant degree of item bias in a test if the samples are large enough, but I doubt that meta-analyses based on different samples from the same populations would substantiate these significant but miniscule item biases found in any one study. Not all that appears as item bias in such studies is actually bias. If blacks and whites differ mainly in *g*, for example, then there will be some reliable degree of group×item interaction in terms of the item's *g* loadings. This effect is seen most clearly in the distinctive profile of the mean black-white differences on various homogeneous subtests of the Wechsler scales (Jensen and Reynolds, 1982). The claim of culture bias can be upheld only if the same item interactions with differences in

ability level fail to appear in comparing culturally homogeneous groups that differ as much in ability levels as do the groups for which cultural bias is claimed. An ideal comparison group for this purpose would consist of sibling pairs (reared together), each member of which is assigned on the basis of test scores to 'high' and 'low' ability groups. (Full siblings differ, on average, 13 to 14 IQ points.) This would insure the groups' perfect equality in cultural background, and any item×group interactions would necessarily be attributed to a difference in ability levels rather than to cultural differences.

Scheuneman asks if test score differences between [black and white] groups are larger than they should be if we knew the 'true' levels of ability. The evidence from predictive validity studies of bias, in which the regressions of criterion performance on test scores are compared in black and white samples, indicates that the common regression line rarely *under*estimates the criterion performance of blacks. Hence, in terms of the ability manifested in the practical performance criteria for which the tests have predictive validity, the test scores do *not* underestimate black ability. As this point is now well established, therefore, it is a misplaced concern to view the tests as the problem. As Lloyd Humphreys (1983) has stated, 'The extent to which minorities are excluded from proportionate participation at all levels in our society is not the result of their lower average test performance. The basic deficit is their performance, on average, in education, industry, and the military' (p. 303).

RACE DIFFERENCES

The study of race differences in intelligence is an acid test case for psychology. Can behavioral scientists research this subject with the same freedom, objectivity, thoroughness, and scientific integrity with which they go about investigating other psychological phenomena? In short, *can* psychology be scientific when it confronts an issue that is steeped in social ideologies? In my attempts at self-analysis this question seems to me to be one of the most basic motivating elements in my involvement with research on the nature of the observed psychological differences among racial groups. In a recent article (Jensen, 1985b) I stated:

> I make no apology for my choice of research topics. I think that my own nominal fields of expertise (educational and differential psychology) would be remiss if they shunned efforts to describe and understand more accurately one of the most perplexing and critical of current problems. Of all the myriad subjects being investigated in the behavioral and social sciences, it seems to me that one of the most easily justified is the black-white statistical disparity in cognitive abilities, with its farreaching educational, economic, and social consequences. Should we not apply the tools of our science to such socially important issues as best we can? The success of such efforts will demonstrate that psychology can actually

> behave as a science in dealing with socially sensitive issues, rather than merely rationalize popular prejudice and social ideology. (p. 258)

Although the study of racial differences constitutes only a small part of my total research efforts, the race theme tends to dominate the overall picture of my activity. The reason is not only that I have probably persisted longer and more systematically than most other researchers who have ventured into this domain, but also that I began by putting the academically tabooed questions 'above board' in a scholarly and factual context that virtually compelled open discussion. So surprising was this that it became an overnight 'media event'. My professional life has never been the same since. Just a few weeks before writing this, my most recent public lecture, at a scientific meeting, on factor analysis and with no reference to race, was picketed by a band of demonstrators. And so it has been ever since 1969.

But that is trivial. It is not what really bothers me. I am much more dismayed by what seems to have become virtually a *de facto* moratorium on research in this area in recent years. The exclusively environmentalist theories of the 1950s that spawned so much psychological and educational research in the 1960s were short-lived in generative power. The research effort fizzled out in the 1970s. Very few of the researchers of that period are still visibly active, at least not in research directed at understanding the nature and causes of the lag of certain minorities in scholastic performance, even though this lag is still proclaimed by educators, government officials, and the media as a persisting and grave problem. Was all the research excitement generated by these questions just a fad of the 1960s? Were not the problems addressed by researchers then just as real today—and just as unsolved? From about 1960 to about 1975 educational research in the United States was dominated by a political philosophy fishing for theories and projects that were consistent with the ideological *Zeitgeist,* and the theories for the most part turned out to be wrong, Policy in public education is at the mercy of politics, and anyone who believes that basic educational research influences politics believes that a sailboat produces the wind. Instead of displaying the cumulative continuity of questions, theory, and investigation that one normally sees in the basic sciences, much of educational research displays merely a varied parade of fads. Those are most favored that are in tune with the prevailing socio-political wind. Why is it that research questions that seemed of vital interest in one decade are abandoned in another? Normally, in science the answer is that either the question was satisfactorily resolved, or it is discovered that the partieular question was scientifically meaningless. But I think it would be hard to argue convincingly that either of these conditions is the case with research on the nature of the black IQ deficit, with all its educational and socioeconomic correlates. If I have done nothing else on this topic, I think I have at least made many psychologists and other social scientists conscious of the inadequacy of our scientific understanding of it. I have shown that the black-white difference on cognitive tests is not a measurement artifact, is not limited to verbal and

scholastic tests, is not associated with any particular classes of informational content of tests but is related more to the complexity of the mental operations required by the items, is not explainable in terms of socio-economic status and is psychometrically distinguishable from social class differences within racial groups, is not explainable in terms of the most popular environmentalist explanations that were scarcely questioned in the 1960s (for example, unequal schooling, teacher expectancy, malnutrition, father absence, verbal deprivation, level of aspiration), and is not dependably or durably reduced to an appreciable degree by any presently known form of educational intervention.

As to the psychometric nature of the difference, I have shown that it is predominantly a difference in *g*, the general factor common to a wide variety of cognitive tasks, rather than a difference in the more specific sources of test score variance associated with any particular informational content, scholastic knowledge, specific acquired skill, or type of test. Hence the difference, whatever its source, cannot be viewed as a superficial phenomenon. Its varying magnitude on diverse tests is related to the tests' *g* loadings (Jensen, 1985a, 1985b). And tests' *g* loadings are not just an artifact of factor analysis. I have discovered that when various tests are rank ordered in terms of their *g* loadings, there are highly significant correlations with the rank order of the tests' correlations with other non-psychometric variables such as heritability of the test scores, degree of assortative mating (spouse correlation), magnitude of parent-child and sibling correlations, degree of inbreeding depression, speed of mental processing in elementary cognitive tasks, and indices based on measurements of the brain's electrical activity (average evoked potential) (Jensen, 1986b).

As to the question of a possible genetic component in the black-white population difference in *g*, since 1969 I have always considered it a reasonable *hypothesis*, for the reasons I have spelled out elsewhere (Jensen, 1973b, 1981a). Based on the total available evidence known to me, I also consider it highly *plausible* that genetic factors are substantially involved in the present black-white population difference. But plausibility falls far short of the status of scientific fact. In science the establishment of a hypothesis as a fact is far more difficult, and must meet far more stringent criteria, often of a highly complex and technical nature, than most non-specialists in the particular branch of science can fully appreciate. There is as yet no empirical 'proof' of this plausible genetic hypothesis of the kind that would be considered as definitive evidence in quantitative genetics. Any other line of evidence that is not strictly *genetic* can only increase or decrease the plausibility of the genetic hypothesis; it cannot lead to certainty in the technical sense that scientific certainty is conventionally established. One (or both) of two kinds of evidence would be required, and neither is likely to be obtained in the foreseeable future: (1) the specific polygenes involved in psychometric *g* would have to be identified and their frequencies in truly random samples of the populations in question would have to be determined; or (2) a true genetic experiment would have to be run in which truly random samples of the two populations are mated in every possible race ×

sex combination and the offsprings are reared in adoptive and non-adoptive homes in every possible race-of-parent × race-of-offspring combination. The offspring would be tested when they reach the age that stable, valid measures of *g* ability can be secured. Such an experiment would permit an analysis of variance attributable to racial genetic and environmental factors. The first alternative is beyond the present technical capability of genetics. Intelligence is conditioned by polygenetic factors and neither the number of such genes nor their chromosomal loci have been discovered; that must await the remote future. The second alternative is technically possible, but practically unfeasible and ethically unacceptable. So the genetic hypothesis will remain untested in any acceptably rigorous manner for some indeterminable length of time, most likely beyond the lifespan of any present-day scientists.

This does not mean, however, that meanwhile there is nothing scientifically worthwhile that present-day psychologists and behavioral scientists can do in this area to advance our knowledge of the nature and correlates of observed racial differences in psychometric abilities. A number of the most feasible and promising avenues for future research on this topic have been well described by Eysenck (1984); there is little I could add. One addition would be a search for genetic pleiotropisms (i.e., two quite distinct phenotypic characteristics connected with one and the same gene) and an examination of their frequencies in different populations. For example, myopia and IQ are positively correlated (both between and within families), which is evidence for pleiotropism, and black and white populations differ in IQ and in the frequency of myopia. A number of other populations would be examined to determine the relationship between mean IQ and frequency of myopia, and the same would be done with other physically measurable pleiotropic characteristics, if they can be identified.

As to how I was drawn into research on racial variation—a question I am frequently asked—I can best answer, in view of the allotted page limitation, by referring to my fairly full accounts of this in my brief autobiography (Jensen, 1974) and in the Preface to my *Genetics and Education* (1973a).

Nichols. This essay attests to the enviable clarity and insight with which Nichols perceives and writes about the issues in the race-IQ controversy. His accurate review of my position and of the typical reactions to it contains nothing I could disagree with. However, one point calls for comment, not because I can fault it, but because it is a point that has troubled me even long before Nichols mentioned it. I feel somewhat chagrined to see someone else point out my own unresolved thoughts so starkly. Nichols states:

> While insisting that racial differences in ability must be understood and explained in scientific terms, Jensen has studiously avoided giving details of how this understanding might contribute to the solution of educational, economic or social problems. In fact, his suggestion for social action, when given at all, is simply to ignore race and to treat each person as an individual. Such a remedy does not depend on knowledge of the cause of

racial differences. Indeed, it is not really a remedy, but a prescription for ignoring the problem.

My reluctance to prescribe stems first of all from my position that it is desirable to maintain a clear distinction between the aims of science *per se*, on the one hand, and its applications in technology, prescription, and policy, on the other. The effectiveness of the latter depends, of course, on commitment, but it also depends crucially on a scientific understanding of the nature of the problem. We see this principle demonstrated in breakthroughs in engineering and medicine. The scientist's job is to find out, rather than to prescribe or to formulate policy. This is because prescription and policy must be based on many other philosophic and economic considerations, such as the proper balance between individual freedom and social welfare, and the allocation of limited resources, for which science can claim no special wisdom. In a democratic society a multitude of interests must come into play in deciding a course of action. Scientific research can only suggest possibilities, and in the light of available theories and evidence scientists can try to predict the probable outcome of a given set of conditions.

Coming specifically to the problem of racial inequality in *g* and its socially important correlates, for example, one could examine birthrates as a function of IQ level within each racial population. Census data suggest that the differential birthrates of low ability and high ability women are less favorable to mean IQ of the black than of the white population. For example, blacks who are college graduates do not even reproduce their own numbers, while the birthrate among blacks who are school dropouts, with no more than an eighth grade education, markedly exceeds the overall average birthrate. The white population shows a significantly less disadvantageous imbalance between low/high ability birthrates. Given the well-established *phenotypic* correlation of about +0.50 between parent and child IQs, it is predictable that if the stated trend in differential birthrates continues, and assuming other conditions remain about the same, the racial IQ gap will gradually increase, thereby magnifying the undesirable conditions mentioned in Nichols' essay. This one problem, if one wishes to think about it at all, gives rise to a branching tree of a great many questions to which scientific answers should inform any suggested remedial prescriptions. Prescriptions can sprout off any branch at any level, and the soundness of the branch can be scientifically examined in competition with prescriptions stemming from other branches. The first question, of course, concerns the validity of the problem as described. Is the evidence adequate to consider it seriously? If the answer is affirmative, one might then ask whether predictably impending natural conditions will soon halt or reverse the trend—a kind of 'spontaneous recovery' —obviating the need for a prescribed remedy. If the most probable answer is negative, then enough scientific knowledge is already at hand for prescribing a remedy with a highly predictable result. The knowledge is simply the phenotypic parent-child correlation of +0.50; the remedy is to control birthrates to reverse the undesirable trend. For any specified degree of control, the results would be

scientifically predictable. But whether such control should be instituted, and the means for doing so, to say nothing of its practical feasibility, involves moral, political, and economic issues that would have to be debated and decided democratically by all interested elements in the whole society. Any prescription with eugenic overtones is almost certainly doomed for the twentieth century, and one can only speculate about conditions in the twenty-first century that may bring about a change in attitudes on this issue.

But the root problem of *g* differences between visibly different populations coexisting in a competitive society gives rise to other branches of the tree, which sprout other remedies. What are the causal factors in the parent-child phenotypic correlation for IQ? Is it more feasible to manipulate these factors than to control birthrates? Are there as yet undiscovered major sources of environmental influence on the development of *g* that, when equalized across racial populations, would wipe out the presently observed *g* difference? Can the demand, reward, and value structures in which *g* has figured so prominently in our technological society be drastically restructured in such a way as to minimize the consequences of real differences in *g* between individuals and between groups? Every one of these questions on which policy decisions might depend can be informed by scientific research. (I can already see I will have to write a book on this eventually.)

Given the present state of our knowledge, and insufficient thought on my part, my own prescription for the time being is to deal as best we can with individual differences and let the statistical group differences fall where they may. Society's general concern with race and other social group differences is not the product of research on these matters, but arises from chauvinist-like attitudes of racial group identity and solidarity in connection with political power and economic interest. It might be termed *meta-racism*. The 'race problem' from that viewpoint is lower in my own hierarchy of values than concern with individual justice and alleviation of individual misfortune. Though it would be blind not to acknowledge the reality of certain statistical differences among populations, I would find it difficult to be the least concerned with any given individual's racial heritage. Perhaps I may be too insensitive on this score, never having felt much sense of racial identity myself.

Flynn. Now and then I am asked by colleagues, students, and journalists: who, in my opinion, are the most respectable critics of my position on the race-IQ issue? The name James R.Flynn is by far the first that comes to mind. His book, *Race, IQ and Jensen* (1980), is a distinguished contribution to the literature on this topic, and, among the critiques I have seen of my position, is virtually in a class by itself for objectivity, thoroughness, and scholarly integrity. My main reservation about the work is that, because of the very nature of the task Flynn has set for himself, there is some constraint on the breadth of evidence he chooses to consider, and I disagree with the weights he assigns to certain items of evidence. I would agree that the rather small body of evidence on which his argument is primarily based—what he terms 'direct evidence' —adds to the

plausibility of the hypothesis that the black-white *g* difference is predominantly attributable to non-genetic factors, perhaps as yet not identified. But weighted in the total picture that a theory must try to accommodate, I do not believe Flynn's case actually tilts the balance against the plausibility of the genetic hypothesis, which, of course, does not exclude the effects of environmental factors either in population differences or individual differences. There are so many peculiarities in the sampling and technical details of the so-called 'direct evidence' racial admixture and adoption studies, which are the basis of Flynn's argument, that I would allow them much less weight in the overall picture than he does. Indirect or circumstantial evidence is not necessarily inferior to direct evidence when the latter is genuinely doubtful. Other behavioral geneticists who have reviewed the same items of 'direct evidence' on which Flynn depends have not found them convincing or given them much weight (for example, Hay, 1985; Loehlin, Lindzey and Spuhler, 1975). I have explicated the reasons for my reservations about the Scarr-Weinberg (1976) cross-racial adoption study elsewhere (Jensen, 1981a, pp. 223–6; 1981b). Hence the title of Flynn's paper—'Jensen's Case Refuted' —seems to me a gross exaggeration. If convincingly true, it would be headline news indeed.

Now Flynn presents another kind of argument not seen in his 1980 book, based on the apparent population changes in raw scores on 'IQ' tests across decades. These rises in test scores are still not understood by psychometricians. They vary in size in different studies, and for different tests, and in different population samples. The inconsistencies are so anomalous that one is forced to wonder if it is possible to obtain a truly random sample of a national population, or equivalent non-random samples, at two widely separated points in time.

While some tests show upward trends over decades, some others show no appreciable changes or even a downward trend. The Scottish National Survey (1949), probably the statistically most impeccable study of intergenerational IQ change ever conducted, showed only a very small rise in Standford-Binet IQ; and the Scholastic Aptitude Test has shown a decline in scores over the past twenty-five years or so. This picture is puzzlingly inconsistent. Only relatively small intergenerational changes, perhaps something less than one-fourth of a standard deviation, could be attributable to genetic factors. Some part of the IQ rise could be caused by the same factors responsible for the significant increase in rate of maturation and adult stature observed in all industrialized nations since the beginning of this century, an effect which has apparently leveled off in the last decade or so. It seems to be largely attributable to improved nutrition and health care.

Part of the problem in interpreting intergenerational changes in IQ is the absence of an absolute scale. The IQ is at best an interval scale, with the meaning of any given IQ being only relative to the population mean at a particular time. Since IQ is not an absolute scale, the meaning of shifts in population means is problematic. Flynn seems to expect that this problematic aspect of the IQ will

detract from the implications of the black-white difference and the plausibility of a genetic hypothesis. I think Nichols' analysis of Flynn's argument is correct.

One of the best established and least arguable facts about IQ and other mental tests is the near constancy across decades in the size of the mean black-white difference measured in standard score units. Recent data on the Armed Services Vocational Aptitude Battery (ASVAB), based on a large national probability sample of youths in the United States, show at least as large a mean black-white difference (about 1.2σ) as was found on similar tests in World War I and World War II. It is noteworthy that these ASVAB results come after two decades of school integration and large-scale compensatory education programs aimed at bettering the intellectual achievements of minorities. The very small, though statistically significant, fluctuations in the black-white difference across certain decades are hardly impressive as compared with the high degree of overall consistency of the difference over the past seventy years.

What our studies of test bias have shown is that black-white IQ differences have the same meaning in terms of external criterion validity as differences of the same magnitude *within* either racial group. What has *not* been demonstrated is that the intergenerational raw-score IQ differences cited by Flynn have equivalent validity in different generations. That is, would a given raw score predict the same absolute level of criterion performance in 1940 as in 1980, assuming random samples of the population at both times? Flynn's data on IQ change on various tests would be much more informative if they could be subjected to some of the methods that are used to assess test bias. Do the across-generation differences show the same absence of test bias as the black-white differences sampled at a given point in time?

Flynn seems to accept the observed differences in test scores across several decades as reflecting true differences in the level of intelligence itself. This would be plausible in the case of relatively small and gradual changes. But a change over the period of three decades (1952–82) of the order of that in the Dutch study—about twenty IQ points—amounts to almost a *reductio ad absurdum* of Flynn's use of these data. It suggests some major, but as yet unknown, artifact. A shift of twenty IQ points for an entire population, if it reflected a corresponding shift in intelligence, or g, with all its well-established correlates, would so drastically change the character of a population as to be absolutely conspicuous. Has any corresponding change in the real-life indicators and correlates of *g* been noticed in the Netherlands between 1952 and 1982? Consider some of the consequences of a twenty-point shift in IQ for a population. We have a fairly clear idea of the practical degree of disability, in school and work, seen in mentally retarded persons with (present) IQs below 70. Such persons are recognized as retarded by their age-peers, their parents, and teachers. In short, persons with an IQ genuinely below 70 are generally recognized as severely handicapped, educationally and occupationally, and they are seldom self-sufficient in the conduct of their personal affairs, often requiring help from relatives or social service agencies. (At present about 2 to 3 per cent of the white

population of the US falls below IQ 70.) Now if the Dutch IQ were really twenty points lower in 1952 than in 1982, and assuming a normal distribution of IQ, we must conclude that there would have been approximately eleven times as many retarded persons, with IQs below 70, in 1952 as in 1982. Is this even remotely plausible? Did the average Hollanders, with IQ 100, who were 25 years of age in 1952, perceive their 50-year-old parents as borderline retarded (IQ between 70 and 80)? At the high end of the IQ scale the approximately top 10 per cent of the population in academic talent who go on to university would have an average IQ of about 125; if the population IQ were raised twenty points, the top 10 per cent going to university would have an average IQ of about 140! Did Dutch professors who were teaching between 1952 and 1982 rejoice over such a great increase in the number of 'geniuses' in their classes? It seems much more plausible that the reported test score increase of twenty points does not reflect a corresponding change in *g* or its real-life correlates, but is rather the result of some artifact not yet identified. These data plausibly appear suspect and call for further investigation.

In his comment on Nichols' paper Flynn notes that the black-white IQ difference does not predict all of the black-white difference in average income. But this should not be surprising. Income does not have a linear regression on IQ throughout its full range. Employability (and the associated income) is partly a threshold phenomenon, so that population means are determined more by the proportions of each population that fall above or below the threshold of minimal job qualification than by the linear regression of job performance or occupational status on ability. More importantly, IQ is only one of many variables related to income. Other behavioral traits and life styles are also related to income, and there are also statistical black-white differences on some of these variables. More research should be done on the correlates of these non-cognitive variables.

It may well be true, as Flynn states, that blacks in America do not 'suffer primarily because of their lack of intelligence.' But still I can imagine that many of them do suffer primarily on this account. White children and adults with poor intelligence, that is, with IQs below, say, 75, certainly suffer considerably in terms of their educational disabilities and their limited job options and poor prospects for advancement. We know there is about a five times greater percentage of the black population (approximately 25 per cent) who in this respect are in the very same boat. Hence, the *g* difference still has tragic consequences. Who could disagree with Flynn that society's struggle to eradicate every vestige of racism must continue? But the causal connection of racism with lowered IQ is still at best an uninvestigated, and perhaps untestable, ad hoc hypothesis. Flynn's plea that 'race, race, race is the primary factor in America's racial problem' cannot explain the conspicuous success of America's Asian minority.

INTELLIGENCE, FACTOR ANALYSIS, AND *g*

The most important theme running through nearly all of my research is the construct of intelligence. It is also perhaps the most important and interesting construct in all of psychology. No other attribute so markedly distinguishes the human species from the rest of the animal kingdom. What is truly amazing is that in the history of psychology many more psychologists have not lavished much more basic research on intelligence than we have seen. Basic research on intelligence has had a checkered history, marked by long periods in the doldrums. (I have written on this history elsewhere [Jensen, 1986a].) Research on strictly the *psychometrics* of intelligence has almost completely dominated the field, with comparatively little interest shown in fathoming the *nature* of intelligence. The amount of literature on the *measurement* of intelligence far outweighs the literature on the *theory* of intelligence. Within just the past decade, however, this strange neglect has begun to change, and there is a rapidly burgeoning new interest in basic research on intelligence. This field of psychology is now beginning to receive its just due.

My study as an educational psychologist gradually led me to the conviction that the *g* factor of our mental tests—whatever *g* is—is by far the most important factor involved in individual and group differences in scholastic achievement. The *g* extracted from a battery of psychometric tests whose contents scarcely resemble anything taught in school is essentially the same *g* as that extracted from a battery of scholastic achievement tests. What more important variable could an educational psychologist concerned with individual and group differences in educability focus on? Working with factor analyses of different collections of diverse tests soon made it obvious, at least to me, that the nature of *g* could not be described or understood in terms of the readily observable, superficial characteristics of the tests in which it loaded. This observation made *g* seem even more fascinating to me. The strong evidence for the substantial heritability of *g* also meant it is not just a psychometric figment, but has its roots in biology. Hence I became increasingly fascinated by *g*. Reviewing all the literature on *g* and related issues was exciting, of course, but it was also unsatisfying, with its plethora of questions and its dearth of scientifically established answers. As I delved further into it, my only lasting regret was that I had not come fully to realize the central importance of *g* much, much earlier in my career. I could have devoted many more years to doing research on it. I greatly doubt that henceforth any other phenomenon, construct, or variable will ever displace *g* in my primary research interest and activity. It would be unfeasible here to summarize my latest views on *g*. I have done this recently in a fairly comprehensive statement (Jensen, 1986b), which concluded as follows:

> An adequate theory of *g* will most probably have to invoke some even more basic level of analysis than is provided by the processing-component sampling theory. It seems likely that continuing effort to achieve a

> scientifically adequate theory of one of the most controversial psychological constructs will force it out of psychology altogether and arrive at an empirically testable formulation in genuinely physiological terms. But this may be the ultimate fate of any truly important construct of psychology. Is it not the ultimate 'psychologists' fallacy' to be satisfied with a psychological explanation of a psychological phenomenos?

Sternberg. This essay, I think, perceives the public Jensen quite clearly and accurately, and I could take exception only to certain details and these only in degree. Naturally, knowing more about myself and my work than is known to Sternberg, I feel that he, too, has somewhat oversimplified the picture of me. But this seems to me inevitable and not essentially objectionable. It is only on those points that Sternberg explicitly expresses his own opinions, rather than in his account of mine, that I find any points of disagreement. Yet I get the impression that Sternberg's few disagreements with me seldom are very fundamental, as if they usually concern style more than substance, and it seems they might mostly evaporate if they were discussed a little further, I think probably because our views of psychology as a natural science are basically much the same.

On the 'Jensen, the *simplifier*' issue, I (Jensen, 1984b) have already had an exchange with Sternberg on this point, It would not be fruitful to expand on it here. We seem to stand at somewhat different points on the 'splitter-lumper' continuum, but not much, at that, compared to the full range seen among all psychologists. The line between *simplification*, which is one of the legitimate aims of science, and overimplification is too subjective and too slippery for a fruitful argument.

The comments on value-free psychology are so vague as to have no teeth. I wish Sternberg had delivered on whatever point he was trying to make by pointing to some actual examples of how my values (or their lack) have led me to 'comparisons that should not be made' or inferences predicated on untrue assumptions. The 'value-free' psychology I would advocate is not free of scientific values, or humanistic moral values, or the value of social responsibility, but I do decry the infestation of psychology, or any science, by political and social ideologies. Ideological contamination of psychological research can only make suspect the claim of psychology to scientific status.

My views on the role of mental speed in intelligence, and the so-called 'oscillation theory', are merely working hypotheses, not ardently held beliefs. I feel no attachment to my working hypotheses; they are merely a means to other ends. I will play with them to see where they lead and discard them, like scaffolding on a building, when they are discredited or no longer useful. I have commented more fully elsewhere (Jensen, 1984c) on my differences with Sternberg regarding the concept of mental speed as a basic factor in intelligence.

The apparently greater complexity of Sternberg's professed view of intelligence than of my view results from his tendency to include in his definition or conception of intelligence many features I would consider merely as correlates

of intelligence or as other variables (personality, motivation, initiative, interests, and the like) that can influence the particular manifestations of a person's intelligence. I prefer a more clear distinction between the construct of intelligence as *g*, on the one hand, and the rich variety of the complex behavioral manifestations of *g*, on the other. Sure, there's more than *g*. In studying intelligence, however, we need not have to imagine that we are studying the whole of human personality and character. Yet who would dispute the 'necessary but not sufficient' property of intelligence in human accomplishment? Sternberg states, 'When we come to think of the predictor—the test—as a better indicator of intelligence than the intelligent performances it is supposed to predict, we are in a bad way.' But if we do not confuse the construct of intelligence with manifest accomplishment, it is quite possible and entirely reasonable that a test could be a better measure of intelligence than the particular performances the test is able to predict with far less than perfect validity.

Brand. Besides its engaging style and numerous quotable epigrams, which make it a delight to read, this essay affords a provocative view of 'IQ' embedded in a remarkably rich context of scientific, social, moral, and philosophic issues. It materially adds to our picture of *g* theory and its many ramifications.

I find only one point in Brand's essay with which I would clearly disagree, not irrevocably, of course, but in terms of my present understanding of the evidence. That is his statement that g, though heritable, cannot have been a fitness character. He bases this conclusion on the fact of the enormous variation seen in human intelligence. I have written on this issue elsewhere (Jensen, 1976a, 1983b). We have two seemingly contradictory types of evidence. On the one hand, there is the consistent evidence for the phenomenon of inbreeding depression on IQ and the finding that the degree of inbreeding depression on various tests is directly related to their *g* loadings. Inbreeding depression depends on the presence of directional genetic dominance, the *g*-enhancing alleles (i.e., alternate forms of a gene) being dominant and the non-enhancing genes being recessive. There is no other genetical explanation for inbreeding depression. It is also known that genetic dominance (for any polygenic trait) has evolved as a result of natural selection favoring the trait in question, and selection (natural or experimental) for a given trait increases dominance. Therefore, the presence of genetic dominance in *g*, as most clearly indicated by inbreeding depression, and the relation of dominance to selection, suggests that *g* is a fitness character affected by natural selection in the course of human evolution. On the other hand, we see large individual differences in *g*. If whatever brain process or processes that underlie the development of *g* ability are enhanced by dominant alleles, and if there has been very strong selection acting on all individuals for many generations, indeed we should expect the dominant alleles gradually to displace the recessive alleles, thereby decreasing the total genetic variance, until eventually the genetic variance is reduced almost to zero, except for statistically small effects of rare mutants that may affect the trait.

The hypothesis that best reconciles these two seemingly contradictory lines of evidence is that selection for *g* has not been especially strong in the distant past and has probably become increasingly relaxed in the last one hundred or so generations. In cooperative social groups selection has less impact on individuals, whose particular characteristics are somewhat buffered against natural selection by protection of the group. All members of a society share in the benefits that arise from the superior capabilities of a small minority of its members. Also, some variation in abilities may have gained an adaptive advantage with the dawn of agriculture and the division of labor that betokened the evolution of civilization. Hence balancing selection, along with the buffering of individual selection (provided there was not a *complete* absence of selection for ability, which would seem most improbable), could very likely result in the considerable degree of heterozygosity that accounts for the genetic variability in intelligence we observe at present.

On another topic, Brand (in his Table 2) lists many often surprising correlates of *g* in the normal population. I am presently preparing a review and meta-analysis of all the known physical correlates of *g.* A few more correlates of *g* could be added to Brand's list: allergies, blood groups, blood serum urate level, leg length (independent of height), basal metabolic rate (in children), the average evoked potential, galvanic skin response, and brain size. Another correlate not listed by Brand is religious affiliation. The causal mechanisms involved in most such correlates of *g* remain a mystery. I have suggested that a reasonable first step in trying to understand the meaning of such correlations is to group them into those I have termed 'adventitious' (i.e., the correlation exists only *between* families) and those termed 'intrinsic' (i.e., the correlation exists *within* families as well as *between* famillies) (Jensen, 1980b, 1984d). Both of these types of correlations indicate the far-reaching manifestations of *g,* by whatever complex chain of causality. But probably only the correlations that qualify as intrinsic will prove to be useful grist in our research aimed at discovering the nature of *g*.

Pellegrino. This uncontentious and straightforwardly expository essay fills in some of the essential background of psychometrics and cognitive psychology that are most germane to my own research. It is a pleasure for me to read something in this vein so lucid and unpolemical. But it leaves me feeling little need for response, besides expressing my appreciation and acknowledging my general agreement with Pellegrino's perception of the topics he treats.

However, I should comment on the next to last paragraph of Pellegrino's chapter, in which he suggests that many persons taking a cognitive ability test may not understand the 'rules of the game' and therefore not bring to bear the particular strategies that make for successful performance. He further suggests that this may also be a source of SES or group differences in test performance. I think that the idea of such strategy factors as a source of differences seems to be largely inconsistent with all the evidence showing extremely high correlations between different populations in the rank order of item difficulty on tests such as Raven's Matrices. Various items call for the induction of different rules for

solution, yet these item differences do not produce differences in the rank order of item difficulty for various social or ethnic groups that differ a standard deviation or so in mean score on the Raven. It seems to me unlikely that this condition would exist if strategy factors were a main source of group differences.

Several years ago I hypothesized that the main determinant of variance in item difficulty (assuming the knowledge content of the items is possessed by all subjects) is the complexity of the cognitive processing demands of the item as reflected in response latency. It has been found that the rank order of mean response latencies to Raven items answered correctly is correlated almost perfectly with item difficulty in terms of percentage failing. A recent study by one of my students (Paul, 1984) examined the relationship between item difficulty (percentage failing) on a simple sentence verification test consisting of items having fourteen levels of complexity in terms of different sentence forms. When the test is taken as an untimed paper-and-pencil test by university students, it is so easy that every subject obtains a perfect score. However, the same items given as a chronometric test, in which response latencies to each item are measured in milliseconds, reveals highly reliable individual differences, as well as marked differences in mean latency between the various sentences. Yet the items are all so extremely simple that the mean response latencies fall in the range of about 650 to 1200 milliseconds, with a response error rate of only 7 per cent. However, when the same items are given as an untimed paper-and-pencil test to third- and fourth-grade school children, the overall error rate is about 17 per cent. The item difficulties for the school children are rank-correlated +0.79 (disattenuated +0.83) with the mean response latencies to the same items by the university students. Yet there is hardly any doubt that the university students understood the ‘rules of the game’ in this very simple sentence verification test, as shown by their short response latencies and the low error rate and the fact that performance was uniformly perfect on the untimed paper-and-pencil form of the test. Thus it appears that whatever features of the items caused mean differences in response latencies among university students were also mainly responsible for the differences in item difficulties among school children. This feature seems to be the complexity of information processing evoked by the item. Students’ subjective ratings of item complexity correlated +0.86 with the item difficulties in the school children and +0.82 with mean item response latencies in the university students. I think that similar applications of chronometric analysis could advance our understanding of the nature of racial and cultural population differences beyond what we are able to learn from traditional psychometric tests alone (see Jensen, 1985a, 1985b; Vernon and Jensen, 1984).

Schönemann. Readers should begin this chapter by reading its final paragraph first. It exposes the real roots of Schönemann’s sophistic diatribe.

Components analysis and factor analysis were invented and developed by the pioneers of differential psychology as a means of dealing with substantive problems in the measurement and analysis of human abilities. The first generation of factor analysts—psychologists such as Spearman, Burt, and

Thurstone—were first of all psychologists, with a primary interest in the structure and nature of individual differences. For them factor analysis was but one methodological means of advancing empirical research and theory in the domain of abilities. But in subsequent generations experts in factor analysis have increasingly become more narrowly specialized. They show little or no interest in psychology, but confine their thinking to the 'pure mathematics' of factor analysis, without reference to any issues of substantive or theoretical importance. For some it is methodology for methodology's sake, isolated from empirical realities, and disdainful of substantive problems and 'dirty data'. Cut off from its origin, which was rooted in the study of human ability, some of the recent esoterica in factor analysis seem like a sterile, self-contained intellectual game, good fun perhaps, but having scarcely more relevance to anything outside itself than the game of chess. Schönemann is impressive as one of the game's grandmasters. The so-called 'factor indeterminacy' problem, which is an old issue recognized in Spearman's time, has, thanks to Schönemann, been revived as probably the most esoteric weapon in the 'IQ controversy'. Out of this factor 'indeterminacy' issue, which few modern factor analysts deem important enough even to mention in comprehensive textbooks on factor analysis, Schönemann has tried to make a mountain out of a molehill. Indeed, there is scarcely a single major modern factor analyst who sees it as more than a molehill, a small one at that. 1 have replied to Schönemann concerning his theme elsewhere (Jensen, 1983c) and will not repeat myself here. My reply to the generic Schönemann, that is, all those who argue that the factors of factor analysis, and *g* in particular, are mere mathematical artifacts without any relation to phenomena independent of psychometrics and factor analysis, is my article, 'The *g* beyond Factor Analysis' (Jensen, 1986b).

Schönemann's arguments about the definition and 'thingness' of intelligence and the meaning of factors were effectively dealt with some forty-six years ago in Chapter 6 (The Metaphysical Status of Factors') of Burt's (1940) *The Factors of the Mind*. When one insists on treating intelligence as a 'thing' rather than as a theory or hypothetical construct intended to generate research, one gets into sophistic arguments that actually seem sophomoric in the light of Burt's chapter. Are mass, gravitation, magnetic field, and potential energy 'things'? Of course not. Why should intelligence, or g, have to be a 'thing' any more than these constructs of physics? What all of Schönemann's harping on factor indeterminacy seems to boil down to is merely a special case of an accepted fact in all empirical science, namely, that all measurement involves some error. This is unavoidable in empirical science, yet all scientific research lives with it and succeeds in advancing our understanding and control of natural phenomena in spite of it. Factor indeterminacy is perceived by psychometricians today as no more of an obstacle to the use of factor analysis in research on intelligence than Olympic runners fear Zeno's Paradox as an obstacle to their reaching the finish line. Schönemann appears to me to view intelligence as a Platonic absolutist. From such a viewpoint all of psychometrics is 'pseudometrics', to use

Schönemann's term. I consider this a nihilistic stance, which, carried to its logical extreme, would reject not only factor analysis but all techniques of measurement and statistical estimation. True, factor scores can only be estimated. But the same can also be said of true scores; and any population parameter can only be estimated from sample statistics. With estimates necessarily go errors of estimate. Does Schönemann's argument imply that the phenomena of interest to psychologists are beyond the grasp of science? There is no more reason to accept this limitation in the case of psychology than in any other science. I think that whatever appeal Schönemann's nihilistic stance may have to some persons merely rides on the back of the current popular antipathy toward 'IQ'. Schönemann's own antipathy on this score comes through loud and clear in his paper.

If *g*, as an estimate of our working definition of the construct of intelligence, is so unacceptable, what would Schönemann propose in its place? The history of science indicates that sheer criticism of a theory or construct carries little force unless it is accompanied by a better formulation. And why use fictitious examples? If Schönemann really has a valid argument, why not use it to show, for example, that *g*, or the largest common factor extracted from different batteries of cognitive tests, is *not* highly similar across the different batteries, or that the mean differences between blacks and whites on various mental tests are *not* more positively related to the tests' *g* loadings than to their loadings on other factors? The reason Schönemann cannot do this is simply that individual differences and the mean differences between populations on a great variety of cognitive tests do not depend in the least on the mathematical machinations demonstrated in his fictitious examples. Until Schöne-mann rolls up his sleeves and tackles the real phenomena of individual and population differences in mental ability, which also show themselves in other realms besides psychometrics, he can hardly be taken seriously. Most educators and employers confronted by Schönemann's sophistry would very likely follow Samuel Johnson, who, on being told of Bishop Berkeley's solipsistic philosophy of subjective idealism, kicked a large stone, exclaiming, 'I refute it thus!'

MENTAL CHRONOMETRY

Galton was the first scientist to put forth the notion of *general ability*, which he conceived in very broad terms. He regarded it as a product of the evolutionary process, and individual differences in it as largely attributable to genetic factors. It was Spearman, however, who invented the methodology for investigating the hypothesis that individual differences in all mental tests, and, indeed, in all kinds of mental performance, reflect differences in a general ability, which accounts for the all-positive intercorrelations among virtually all tasks of a cognitive nature. The demonstration of a general, or *g*, factor in any matrix of correlations among various mental tests was seen as evidence supporting the hypothesis of a general ability. In trying to fathom the nature of *g*, Spearman depended upon

trying to characterize the common features of those tests, among one hundred or so diverse tests, that factor analysis revealed as having the largest loadings on the *g* factor. By this criterion Spearman's characterization of *g* as 'the eduction of relations and correlates' was correct and is still valid, as far as it goes. But it caused most psychologists to view *g* primarily as reasoning ability involving 'higher thought processes' and strategies for problem-solving, particularly of a scholastic nature, because *g* was also found to be substantially correlated with indices of scholastic achievement. The study of *g* exclusively in terms of the tests that were most highly *g*-loaded lost sight of the many other tests that were also loaded on *g*, albeit not very highly, but did not seem to involve complex reasoning. Early in his research Spearman claimed that even pitch discrimination and other relatively simple sensory tasks have some small loading on *g*, as though there was no point on the whole continuum of task complexity that showed a break in the smooth distribution of *g* loadings. Tasks' loadings on *g* appear as a smooth continuum, ranging from very near zero up to nearly the reliability of certain tests, provided the tests are obtained from an unrestricted sample of the general population.

I was intrigued by the fact that tests that did not seem to be characterized by relation eduction or other forms of complex reasoning nevertheless still had some significant loading on *g*, and therefore correlated with highly g-loaded tests that they did not superificially resemble in the least. It suggested that the Spearman characterization of *g* was too narrow, and that *g* might really be closer to the broader Galtonian notion of general ability. How far down on the continuum of task complexity could the same *g* that loads highly on such complex reasoning tests as Raven's Matrices still be found? Might not a battery of such simple, but g-loaded, tasks be able to reveal something about the nature of *g* that so far psychologists had not discerned in their use of complex tests? Obviously, if tests were to be made so very simple that every normal person could perform the tasks, the only means of measuring individual differences would be to measure response latency, or reaction time.

Such were my thoughts in the early 1970s. I knew that Galton had used reaction time (RT) tests and that his followers, such as James McKeen Cattell and his student Clark Wissler, had carried on Galton's work. These early studies found practically no relationship between RT (or other simple functions) and such limited criteria of mental ability as college grades. These studies were incredibly weak. In view of the poor reliability of the RT measurements and the criteria with which they were correlated, and the restricted range of general ability in the samples tested, it was no wonder that correlations between RT and 'intelligence' were close to zero. They could hardly have been otherwise under these conditions. The few other old studies of RT and intelligence that I found in the literature were scarcely better, although a number of studies had shown that the mentally retarded had slower RTs than normals. But I was more interested in variation within the normal range of IQs, and the evidence on RT and IQ in this range was not only inconclusive but almost non-existent. One rather obscure

study by Roth (1964), that I had found reference to in an article by Eysenck (1967), caught my attention. Roth had suggested a technique for measuring RT that seemed to make sense theoretically and yielded promising results. Roth's method was based on Hick's law—that RT is a linearly increasing function of the logarithm of the number of choice alternatives among which the given reaction stimulus is presented. Roth interpreted the slope of this function as a measure of the speed of information processing (in milliseconds per bit of information, where a bit is the binary logarithm of the number of alternatives in the array of potential reaction stimuli). He reported a negative correlation between RT slope and psychometric intelligence.

I devised a similar apparatus, but used a procedure that divided the total time for the subject's reaction between (1) RT *per se* (the interval between onset of the reaction stimulus (a light going on) and the subject's removing his finger from a 'home' button and (2) movement time, MT (the interval between the release of the 'home' button and pressing a button adjacent to the reaction stimulus, turning it off). (Detailed descriptions of the apparatus and procedure for the Hick paradigm and other techniques used in my RT research can be found in Jensen, 1985c.) I still think it is very important to separate RT (also called decision time) from MT in all studies of the speed of information processing, because RT and MT are not highly correlated, even when the correlation is disattenuated, and so lumping them together confounds the underlying latent variables, a highly undesirable condition when we are studying the correlations between mental speed in elementary cognitive tasks and scores on psychometric tests of ability.

From my standpoint there is nothing especially important or interesting about RT *per se*. I see it merely as a technique for studying individual differences in cognitive tasks that are so simple and elementary (hence called elementary cognitive tasks or ECTs) that, except for very young children and the profoundly retarded, the only possible reliable measure of individual differences is latency of response. Even bright university students show highly reliable individual differences in ECTs that are so simple that their response latencies, or RTs, are less than 1 second. The amazing thing is that these very brief RTs to a variety of ECTs are significantly correlated (negatively) with scores on complex psychometric tests given under non-speeded conditions. This finding means that complex culture-loaded tests, such as the Wechsler scales, are actually measuring individual differences in something other than the knowledge content of the tests or particular complex skills and strategies for solving problems considered to be of an 'intellectual', if not entirely scholastic, nature.

A question of major theoretical interest is how much of the variance in the psychometric *g* represented in our standard IQ tests is accountable in terms of ECTs. What is the nature of the ECTs that are correlated with *g*? And what is the upper limit of the *g* correlation that can be found for ECTs at a given level of complexity? Must ECTs involve higher-level strategies, or meta-processes, in order to show a substantial correlation with the *g* of complex psychometric tests?

An interesting and theoretically important working hypothesis is that individual differences in psychometric *g* derived from non-speeded tests reflect differences in the speed or efficiency of mental processing of information, and that the same differences in speed are measurable in tasks making such simple cognitive demands that correct responses have mean RTs around 1 second or less.

In my laboratory, using several different elementary tasks in combination, my co-workers and I have found replicable correlations between the composite RTs and scores on psychometric tests (for example, Raven Matrices, Wechsler, Terman Concept Mastery, Armed Services Vocational Aptitude Battery) that are almost as high as the correlations between different psychometric tests in the same study samples. To explore the generality of the phenomenon, we have looked for correlations in a wide range of samples, from the severely retarded to the academically gifted, and have found similar relationships (taking into account group differences in reliability and restriction of range) in the various groups at every level of IQ. Single RT tasks have a correlation ceiling of about 0. 50, and correlations are more typically around 0.30. I think this ceiling is due to the large amount of task-specific variance in any one RT paradigm. In this respect an RT task behaves more like a single test item than like a test composed of various items, permitting item specificities to 'average out' in the total score. Any particular RT task, such as the Hick paradigm or the Sternberg memory-scan paradigm, is extremely homogeneous as compared with typical psychometric tests, and therefore has much more specificity. The solution is to employ a battery of diverse RT tasks. A part of my present research effort is directed at finding a number of RT tasks that, in combination, will yield maximal correlations with *g*. This problem itself involves questions of theoretical importance that cannot be adequately explicated here. For example, does the larger correlation with *g* produced by a battery of RT tasks (as compared with any single RT task) depend on the tasks' tapping a number of *different* hypothesized elementary cognitive processes (for example, stimulus encoding, discrimination, choice, short-term or long-term memory retrieval, rotation of mental images), or does it depend on merely varying the tasks sufficiently to 'average out' the task-specific variance, even without increasing the number of different hypothesized *g*-related cognitive processes?

Our simplest working hypothesis is that any and every ECT involves the same *g* to some extent (as do all items of psychometric tests), and it does not matter which particular ECTs enter into a battery, as long as there are enough of them to 'average out' their specificities. This hypothesis, if substantiated, would be a further demonstration of Spearman's 'theorem of the indifference of the indicator' of *g*. It is a crucial hypothesis for the 'cognitive components' theory of *g*—the idea that *g* variance depends on variance in a number of distinct cognitive processes that are sampled by psychometric tests. If measures of these distinct processes are themselves highly intercorrelated (after correction for attenuation), showing, when factor analyzed, much the same *g* as the *g* of psychometric tests, the search for the nature of *g* would have to be extended to a more basic level

than that envisaged in componential theories. It is toward this fundamental issue, I think, that the RT research by me and many others is headed.

Eysenck. This essay views my RT research in the broad Galtonian context for the study of mental ability that lends RT its theoretical interest, the full importance of which, I think, has not yet been perceived by many contemporary psychologists. Eysenck fully appreciates the broad theoretical implications of this line of investigation, and his chapter is an excellent summary of the key issues at present.

The most basic hypothesis to which my RT research is addressed is well stated by Eysenck: '…there is a central core to IQ tests which is quite independent of reasoning, judgment, problem-solving, learning, comprehension, memory, etc.' This hypothesis, in my opinion, is presently more strongly supported by a number of lines of evidence than the contrary hypothesis that *g* reflects only a sampling of various tasks of reasoning, problem-solving, etc., or a sampling of the hypothesized cognitive processes and meta-processes that are hypothesized to enter into such tasks. The hypothesized 'central core' that Eysenck refers to is probably not even describable in terms of psychological or cognitive concepts. Such concepts, however, are legitimate and probably essential in attempting to describe the varied manifestations of individual differences in the 'central core'.

Eysenck correctly notes the as yet highly tentative nature of my theoretical formulation of the connection between RT and *g*, and he points up an important theoretical gap (which is also an empirical gap), namely, the relationship of RT to Level I ability. The limitation of short-term memory capacity, or so-called working memory, is a part of my hypothesis concerning the mechanism through which speed of information processing becomes a fundamental variable in *g*, as clearly explained by Eysenck. But what is the relationship between individual differences in the speed of processing, as indicated by choice RT, and the capacity of working memory, as indicated by forward digit span? From my notion of Level I/Level II, it was my hunch that RT and memory span would be uncorrelated, and that individual differences in working memory capacity constitute only a relatively small part of the variance in *g* as compared with speed of mental processing. The only study I did on this, with fifty university students, using the Hick paradigm for RT, the Raven (as a measure of *g* or Level II), and forward digit span (as a measure of short-term memory capacity), yielded the following correlations (asterisk indicates significance at the .05 level, two-tailed):

Digit span	×	Raven, r=+0.22
RT intercept	×	Raven, r=+0.15(+.0.03)
RT slope	×	Raven, r=−0.41*(−0.39*)
RT intercept	×	Digit span, r=+0.16(+0.15)
RT slope	×	Digit span, r=−0.04(+0.01)
RT intercept	×	RT slope, r=−0.29*

The correlation between RT intercept and slope is negative completely due to the artifact of correlated measurement error; the very same errors of measurement have *opposite* effects on the magnitudes of intercept and slope. Therefore, a more accurate correlation between either intercept or slope with an outside variable can be obtained by partialling out the effect of either variable (intercept or slope) from the correlation of the other variable with the Raven or with digit span. These partial correlations are shown in parentheses. The relative sizes of these correlations appear quite consistent with my hypothesis. If RT slope measures speed of information processing (greater slope=slower speed), it should be more correlated with the Raven (r=−.39) than with digit span (r =+0.01). No attempt has been made to replicate these results. But they seem questionable because few other studies have found such a high correlation between RT slope (in the Hick paradigm) and any test of g, and several studies have found near-zero correlations. Yet *groups* that clearly differ in *g* quite consistently differ in RT slope in the theoretically predicted direction. (See the further discussion of this point under my comments on *Carroll*.)

Clearly, we need further studies of the relationship of RT parameters to working memory capacity. In such studies I think it important to measure working memory capacity as a broader trait than merely forward digit span. First principal component factor scores should be derived from a battery of memory span tests in which the materials are varied, so as to minimize task specificity, using not only digits, but letters, simple words, colors, forms, color-forms, symbols, pictures of familiar objects, Knox cubes, pitch patterns, and the like. It would be important to determine the degree of correlation between the largest common factor in such a battery of tests of short-term memory capacity and the *g* factor of complex tests such as the Raven and the Wechsler. If the largest common factor in such a battery of simple memory tests turned out to be much the same as the *g* of intelligence tests, it would strongly suggest that the Level I/ Level II distinction is largely an artifact of the large amount of specificity in the few measures of Level I we have used. In short, we are not at all certain of the degree of independence of individual differences in psychometric *g* and in working memory capacity when it is measured as the largest common factor of a number of diverse tests of memory capacity.

Carroll. Despite what seems to me its unrelieved negative tone, I find this hard-hitting critique most useful for presenting what is perhaps the strongest case that can possibly be made against the research and theoretical implications derived from just one of the RT paradigms that has been used in my investigations of the hypothesis that speed of information processing is importantly and causally related to psychometric *g*. Certainly, few, if any, other experts in this field are technically more qualified for executing this ‘onerous task’ than Professor Carroll. His renown as a methodologist and formidable critic, in addition to his encyclopedic knowledge of the literature on information processing and intelligence (for example, Carroll, 1980), compel our most thoughtful consideration of the key points of his critique.

However, its critical focus exclusively on the Hick paradigm, which is only one of the several RT paradigms investigated in my laboratory in recent years, creates, I think, an unduly narrow view of what my co-workers and I have been doing. It was partly because of certain limitations of the Hick paradigm and my dissatisfaction with the puzzling inconsistencies in some of its results across different subject samples that I have added other, more complex, RT paradigms to our battery of techniques. The results from the Hick paradigm become more meaningful when viewed in relation to the other RT paradigms. For example, the suggestive but often inconsistent increase in correlations between RT and *g* as a function of increasing task complexity (i.e., number of bits) shows up more strongly and consistently in our more complex RT paradigms, which lends credence to the same but weaker trend in the Hick paradigm. (Note Carroll's Table 1. The correlations of RT with Raven as a function of 0, 1, 2, and 3 bits respectively, averaged over all samples, are −.19, −.23, −.24, and −.27 respectively; these correlations have a linear correlation with bits of -0.98 [$p<.01$] and hence the overall trend of these data is not inconsistent with the hypothesis that the correlation of RT with *g* increases as a function of bits.) By focusing on just the inconsistencies in the data, as in his Table 1, Carroll loses sight of the overall picture.

Nearly all the critical points raised by Carroll are of such a nature that a proper response to them depends on more explication of technical matters and tabular presentation of results from a number of studies (some not available to Carroll at the time of writing), along with meta-analyses of the means, intercepts, slopes, correlations, and other statistics from the various study samples, than is feasible in the present chapter. I have done this kind of summary meta-analysis of all our results on the Hick paradigm in a highly detailed chapter of a book concerned entirely with research on RT and intelligence (Jensen, in press). For example, it is now possible to examine Hick parameters (mean RT, intercept, slope) and their correlations with 'IQ' based on twenty-four independent samples totalling more than 1500 subjects. Unfortunately, it is not feasible to report the analyses of this material here in the detail required for a proper response to Carroll's criticisms, most of which may appear rather deflated when we can see in perspective the whole forest as well as the trees. The critique by Longstreth, which is cited approvingly by Carroll, is even more strikingly diminished by critical examination. The Longstreth article, in fact, is an item in evidence against the all too common presumption that a critique is much less liable to faultiness than the things it criticizes. A detailed reply to Longstreth's critique has been submitted to the journal in which it appeared.

Carroll (and also Longstreth) apparently choose to ignore all the data on mean differences in RT parameters (means, intercepts, and slopes) between groups that differ in average level of intelligence. These group mean differences can also be used to test the hypothesized relationships between RT parameters and intelligence. Group mean differences have the advantage that measurement error tends to be averaged out in the mean. On the other hand, when the theoretically

expected correlation is moderate and the groups are of moderate size (the typical *N* in our studies is 50), there is considerable sampling error in any within-group correlations. For example, in any one of the comparison populations with a restricted ability range from which a study group is sampled, if the true correlation between an RT parameter and IQ is, say, 0.30, then, for samples of *N*=50, 68 per cent of the obtained within-group correlations can be expected to fall in the range of correlations between 0.17 and 0.43, and 99 per cent will fall between .05 and 0.55. This variability in obtained correlations makes it more important to look at meta-analyses of correlations from numerous studies (the 'forest') than just at each single correlation (the 'trees'). Other indicators of relationship, such as mean differences between various criterion groups that differ in psychometric *g*, should also be considered. It is rare to find group differences in any RT parameters that are inconsistent with their hypothesized relationships to *g*, as I show in Jensen (in press). For example, groups differing in mean IQ also show highly significant differences in RT slope (in the Hick paradigm) in the theoretically predicted direction with overwhelming consistency. Should we completely ignore such findings or dismiss them because some of the within-group correlations between slope and IQ are non-significant?

On at least one point Carroll gives the impression that his interpretation of the data is at odds with mine. He states that mean RT could be only a 'proxy' variable for RTSD (i.e., the standard deviation of RTs over trials, as a measure of intra-individual variability), and he urges 'caution in thinking of mean RT as a variable unaffected by intra-individual variability, as Jensen appears to do….' But I myself have made precisely the same point: Theoretically, too, variability of RTs would seem to have priority over the average speed of RTs…. The average speed of RT can be seen as a consequence of variability of RT more easily than the reverse relationship' (Jensen, 1982b, p. 103).

Although I cannot fully present the basis of my conclusions here, and readers must be referred elsewhere for this (Jensen, in press), I will nonetheless mention the several points on which I have some disagreement with Carroll.

I doubt that spatial ability *per se* is important in the Hick performance or its correlation with IQ, partly because the Raven is a very weak measure of spatial ability and because other non-spatial tests are correlated with Hick parameters. RT slope correlates less with the WISC-R subtests most likely to have a spatial component (Mazes, Block Design, Object Assembly) than with verbal tests (Hemmelgarn and Kehle, 1984). The average correlation of RT slope with the Vocabulary, Information, Similarities, and Comprehension subtests was −0.26, as compared with an average correlation of –0.13 for Block Design, Mazes, and Object Assembly. The twelve WISC-R subtests' correlations with RT slope were correlated +0.80 with the subtests' *g* loadings, suggesting that RT reflects *g* more than a spatial factor. Some spatial tests are also highly g-loaded, and *g* variance would need to be statistically controlled in any study aimed at the hypothesis that spatial ability is importantly reflected in RT parameters. Other RT paradigms,

too, have shown that the magnitude of correlations between RT and various psychometric tests is directly related to the tests' *g* loadings (Jensen, 1986b).

It seems highly improbable that speed-accuracy trade-off could account for the RT correlation with IQ, since higher IQ is associated both with *lower error rate* and with *shorter RT*. Longstreth's contrary and implausible speculations on this issue are totally without empirical support. I have no argument with the attentional hypothesis of intra-individual RT variability, but fluctuations in attention may only be a reflection of the same underlying process involved in RT variability. Invoking attention as an explanatory construct in this context does not seem to get us anywhere. Fluctuations in attention are no better understood than variability in RT.

I differ with Carroll's opinion that the study of intelligence is 'better approached through analysis of the tasks actually employed in cognitive ability tests themselves' than through measures of mental processing speed derived from specially contrived laboratory tasks that have little or no resemblance to the traditional ability tests. The most important finding I have seen come out of the type of research advocated by Carroll is that a general speed-of-processing factor common to a number of different cognitive processing components that enter into complex ability tests, such as verbal and figural analogies, shows a much more substantial correlation with psychometric *g* than do any of the processing components independent of their largest common factor, which appears to be speed of mental processing. This is one of the main conclusions arising from Sternberg's componential analysis of analogical reasoning, a type of task commonly used in traditional intelligence tests. When the amounts of time required for execution of each of the several component processes in the analogies tasks are entered into a multiple regression to predict IQ, or psychometric *g*, what is found? In Sternberg's (1979a) words:

> Information-processing analyses of a variety of tasks have revealed that the 'regression constant' is often the individual differences parameter most highly correlated with scores on general intelligence tests. This constant measures variation that is constant across all of the item or task manipulations that are analyzed via multiple regression. The regression constant seems to bear at least some parallels to the general factor. (p. 24)

Referring to the same point elsewhere, Sternberg (1979b) says this about the 'regression constant': '…we can feel pleased to be rediscovering Spearman's *g* in information processing terms.' Therefore, it seems to me that the speed factor common to the various component processes involved in complex cognitive tests merits study in its own right. It can probably be made more accessible to chronometric analysis by means of comparatively simple laboratory tasks specially devised to measure particular facets of processing speed. But an even more basic reason that RT tasks with very little resemblance to psychometric tests interest me is the very fact of their little resemblance. This allows the

correlation they have with psychometric factors to extend the meaning of those factors beyond the confines of psychometric tests. To find that the common factor of a number of simple chronometric tasks that bear no surface resemblance to IQ tests is correlated with the *g* of IQ tests is, at least to me, a much more pregnant phenomenon, scientifically, than a demonstration that chronometrically derived components of IQ test items are correlated with the *g* factor derived from the very same or highly similar tests. Both types of investigation, of course, are necessary for an adequate account of the role of mental speed in cognitive performance. Speed itself is probably a derivative behavioral phenomenon resulting from some more fundamental neural processes in the brain, which at present have been couched in such embryonic constructs as neural oscillation, error tendencies, or 'noise', in the neural transmission of information. I agree with Carroll that these notions are highly speculative at this time and that a much more detailed network of consistent empirical findings will be needed before we can get a scientific handle on such theoretical speculations. At this stage there are too many possible hypotheses, but there is not yet nearly enough empirical knowledge to evaluate them or to constrain our speculations in a scientifically productive way.

Carroll's point, that the fact that a number of variables all show substantial positive loadings on the unrotated first principal component (or first principal factor) does not necessarily mean that all of the variables are positively intercorrelated, is unarguably correct. I must agree that the first principal component presented in my 1979 paper, referred to by Carroll, was a mistake, because of its implication that the Raven (a marker for psychometric *g*) and Concept Mastery Test (being loaded +0.73 and +0.57 respectively on this component on which a number of RT variables were also very substantially loaded) were highly correlated with RT. (The reflected zero-order correlation between Raven scores and RT slope was +0.410; between CMT and slope, +.002; between Raven and CMT, +0.402.) Hence, I now regard it as far preferable, indeed essential, to represent the general factor, in the sense of Spearman's g, by means of a hierarchical factor analysis, for which I have found the Schmid-Leiman (1957) method the most useful.Usually, the first unrotated principal component (or factor) and the hierarchical general factor are extremely similar, but this is not a mathematical necessity, and so the hierarchical analysis (for example, Schmid-Leiman) yields *g* loadings that cannot give a misleading impression of the true generality of the *g* factor in the correlations among all the variables in the matrix. But the particular components analysis that Carroll has rightly criticized for the reason I have just mentioned also has a more serious fault that no one, to my knowledge, has yet pointed out—more serious because it would contaminate any type of factor analysis, including a hierarchical factor analysis. I refer to the inclusion of RT intercept and RT slope together in the same factor analysis. I did not know it at the time, and apparently scarcely anyone else did, but I now realize this is a serious mistake. I mention it to warn others. The correlation between intercept and slope is largely artifact due to their negatively

correlated errors of measurement. The only way to get around this, if for any reason intercept and slope must be entered into the same factor analysis, is to derive each of these parameters from experimentally independent sets of data, so that their measurement errors will have zero correlation with one another. In general I would now urge the same treatment for any other parameters derived from one and the same set of RT measurements.

Fortunately, we know from the history of science that if research along a particular line is carried on long enough and assiduously enough by a number of investigators, the dross noted by critics is gradually filtered out and forgotten, leaving, one hopes, enough ore to repay the effort of the research. From this viewpoint I am probably more hopeful than Carroll about the eventual value of my use of RT measurements in the study of individual differences in intelligence. Time will tell.

EDUCATIONAL AND SOCIAL IMPLICATIONS

More than ten years ago, while spending a summer in London, I was requested by the editors of the *Oxford Review of Education* to write an article on what I considered the broad educational and social implications of our present knowledge of differential psychology. This article, entitled 'The Price of Inequality' (Jensen, 1975b) is my only attempt so far to concentrate on this broad moral and philosophic aspect of our study of human differences. Although neither Bereiter nor Havender makes any reference to this article, virtually all the thoughts expressed in it are brilliantly and profoundly amplified in their own essays, which also point up a number of important insights that had not occurred to me. The ideas expressed in these chapters are essentially so concordant and intermeshed that I feel no need to comment on them separately.

I find myself in agreement with everything they say, while recognizing that much of what can be said in this particular realm at present is necessarily based on opinion and philosophic outlook. Both Bereiter and Havender seem to hold out more hope for aptitude-by-instruction interaction as a partial solution to the problem of individual differences in scholastic achievement. I would agree that the search for useful interactions should not be abandoned, but I have seen little so far that would make me optimistic on this score. I have begun to ask *why* it is that interactions, at least with respect to *g*, the single greatest source of variance in scholastic performance, have been so hard to discover or to demonstrate. Perhaps the polygenic and polyenvironmental model of a multitude of small additive effects *is* the most realistic explanation of the sources of individual differences in *g*, and therefore it is virtually impossible by any feasible environmental means to manipulate individual differences (and *ipso facto* group differences) in *g*. After all, it should not be forgotten that about half of the population variance in *g* and in scholastic achievement exists *within* families (i.e., sibships), and this half of the variance is entirely attributable to polygenic and micro-environmental factors. The almost negligible correlations between the

IQs of *nominal* siblings (i.e., unrelated children reared together by adoptive parents) suggest that most of the non-genetic variance in IQ is of the within-family micro-environmental variety. Could it be that the biological underpinnings of *g* have evolved so as to minimize interaction with different environmental contingencies in order to maximize the generality of *g*? The very *generality* of ability, which seems to be a distinguishing feature of Homo sapiens, may be an important product of the evolutionary process, serving to safeguard the behavioral capacities of the species from being too much at the mercy of any particular environmental happenstance. My 'best guess' at present is that while there are important ways in which education can be improved, to the great benefit of individuals and society, an appreciable increase in intelligence, in the sense of *g*, will not be one of the effects. It seems to me most likely that *g* variance will prove to be manipulable to any practically significant degree only by some essentially biological means, such as genetic selection or direct intervention at some point in the causal chain between genes and behavior. This is scarcely on the horizon at present. But there are still many other possibilities for improving the outcomes of education, which depend on other important variables besides *g*. The fact that whole schools, communities, and nations show greater average differences in their educational products than can be attributed to their differences in *g* indicates that other factors must also play an important part —educational values, the work ethic, parental support, motivation, time on task, and efficiency of instructional methods, to mention a few.

As for the various group differences that exist in any large national population, especially those differences that are consequential for schooling and occupations, the only reasonable, just, and moral stance I know at present is the one that is so well put in the final paragraph of Havender's essay. Read it again.

It has conspicuously fallen to the lot of this generation of behavioral scientists to seek an understanding of the human variation that must inevitably challenge the wisdom of every caring society. Already, before this has gone to press, I am looking forward hopefully to the future time, perhaps not too distant, when the controversies discussed in this book will all seem like 'ancient history', the authentically important questions finally yielding to sufficient facts to enable a scientifically worthy consensus.

REFERENCES

Agrawal, N., Sinha, S.N. and Jensen, A.R. (1984) 'Effects of inbreeding on Raven Matrices', *Behavior Genetics,* **14,** pp. 579–85.

Burt, C. (1940) *The Factors of the Mind,* London, University of London Press.

Carroll, J.B. (1980) *Individual Difference Relations in Psychometric and Experimental Cognitive Tasks,* Chapel Hill, N.C, L.L. Thurstone Psychometric Laboratory, University of North Carolina.

Eysenck, H.J. (1967) 'Intelligence assessment: A theoretical and experimental approach', *British Journal of Educational Psychology,* **37,** pp. 81–98.

Eysenck, H.J. (1984) 'The effect of race on human abilities and mental test scores,' in Reynolds, C.R. and Brown, R.T. (Eds), *Perspectives on Bias in Mental Testing*, New York, Plenum, pp. 249–91.

Flynn, J.R. (1980) *Race, IQ, and Jensen*, London, Routledge and Kegan Paul.

Goddard, H.H. (1913) 'The Binet tests in relation to immigration', *Journal of Psycho-Asthenics*, **18,** pp. 105–7.

Goddard, H.H. (1917) 'Mental tests and the immigrant', *Journal of Delinquency*, **2,** pp. 243–77.

Hay, D.A. (1985) *Essentials of Behavior Genetics*, Melbourne, Blackwell.

Hemmelgarn, T.E. and Kehle, T.J. (1984) 'The relationship between reaction time and intelligence in children', *School Psychology International*, **5,** pp. 77–84.

Humphreys, L.G. (1983) 'Review of "Ability Testing"' (Ed. by Wigdor, A.K. and Garner, W.R.), *American Scientist*, 71, pp. 302–3.

Hunter, J.E., Schmidt, F.L. and Rauschenberger, J. (1984) 'Methodolgical, statistical, and ethical issues in the study of bias in psychological tests', in Reynolds, C.R. and Brown, R.T. (Eds), *Perspectives on Bias in Mental Testing*, New York, Plenum, pp. 41–99.

Jensen, A.R. (1961) 'Learning abilities in Mexican-American and Anglo-American children', *California Journal of Educational Research*, **12,** pp. 147–59.

Jensen, A.R. (1967) 'Estimation of the limits of heritability of traits by comparison of monozygotic and dizygotic twins', *Proceedings of the National Academy of Science*, **58,** pp. 149–56.

Jensen, A.R. (1968a) 'Social class, race and genetics: Implications for education', *American Educational Research Journal*, **5,** pp. 1–42.

Jensen, A.R. (1968b) 'Another look at culture-fair testing', *Western Regional Conference on Testing Problems, Proceedings for 1968: Measurement for Educational Planning*, Berkeley, Calif., Educational Testing Service, Western Office; reprinted in Hellmuth, J. (Ed.) (1970), *Disadvantaged Child, Vol 3, Compensatory Education: A National Debate*, New York, Brunner/Mazel, pp. 53–101.

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

Jensen, A.R. (1970) 'Selection of minority students in higher education', *Toledo Law Review*, Spring-Summer, Nos 2 and 3, pp. 304–457.

Jensen, A.R. (1973a) *Genetics and Education*, London, Methuen.

Jensen, A.R. (1973b) *Educability and Group Differences*, London, Methuen.

Jensen, A.R. (1973c) 'Level I and Level II abilities in three ethnic groups', *American Educational Research Journal*, **4,** pp. 263–76.

Jensen, A.R. (1974) 'What is the question? What is the evidence?' (autobiography), in Krawiec, T.S. (Ed.), *The Psychologists*, Vol. 2, New York, Oxford University Press, pp. 203–44.

Jensen, A.R. (1975a) 'A theoretical note on sex linkage and race differences in spatial ability', *Behavior Genetics*, **5,** pp. 151–64.

Jensen, A.R. (1975b) 'The price of inequality', *Oxford Review of Education*, **1, 1,** pp. 13–25.

Jensen, A.R. (1976a) 'Heritability of IQ', in 'Letter to the Editor', *Science*, **194,** pp. 6–14.

Jensen, A.R. (1976b) 'The problem of genotype-environment correlation in the estimation of heritability from monozygotic and dizygotic twins', *Acta Geneticae Medicae et Gemellologiae*, 25, pp. 86–99.

Jensen, A.R. (1979) '*g:* Outmoded theory or unconquered frontier?', *Creative Science and Technology*, 2, pp. 16–29.
Jensen, A.R. (1980a) *Bias in Mental Testing*, New York, The Free Press.
Jensen, A.R. (1980b) 'Uses of sibling data in educational and psychological research', *American Educational Research Journal*, 17, pp. 153–70.
Jensen, A.R. (1980c) 'Level I and Level II abilities in Asian, white, and black children', *Intelligence*, 4, pp. 41–9.
Jensen, A.R. (1981a) *Straight Talk about Mental Tests*, New York, The Free Press.
Jensen, A.R. (1981b) 'Obstacles, problems, and pitfalls in differential psychology', in Scarr, S. (Ed.), *Race, Social Class, and Individual Differences in IQ*, Hillsdale, N.J., Erlbaum, pp. 483–514.
Jensen, A.R. (1982a) 'The chronometry of intelligence', in Sternberg, R.J. (Ed.), *Advances in the Psychology of Human Intelligence*, Vol. 1, Hillsdale, N.J., Erlbaum, pp. 255–310.
Jensen, A.R. (1982b) 'Reaction time and psychometric *g*', in Eysenck, H.J. (Ed.), *A Model for Intelligence*, New York, Springer-Verlag, pp. 93–132.
Jensen, A.R. (1983a) 'Sir Cyril Burt: A personal recollection', *Association of Educational Psychologists Journal*, 6, pp. 13–20.
Jensen, A.R. (1983b) 'Effects of inbreeding on mental-ability factors', *Personality and Individual Differences*, **4,** pp. 71–87.
Jensen, A.R. (1983c) 'The definition of intelligence and factor-score indeterminacy', *The Behavioral and Brain Sciences*, **6,** pp. 313–15.
Jensen, A.R. (1984a) 'Test bias: Concepts and criticisms', in Reynolds, C.R. and Brown R.T. (Eds), *Perspectives on Bias in Mental Testing*, New York, Plenum, pp. 507–86.
Jensen, A.R. (1984b) 'Jensen oversimplified: A reply to Sternberg', *Journal of Social and Biological Structures*, **7,** pp. 127–30.
Jensen, A.R. (1984c) 'Mental speed and levels of analysis', *The Behavioral and Brain Sciences*, **7,** pp. 295–6.
Jensen, A.R. (1984d) 'Sociobiology and differential psychology: The arduous climb from plausibility to proof', in Royce, J.R. and Mos, L.P. (Eds), *Annuals of Theoretical Psychology*, Vol. 2, New York, Plenum, pp. 59–88.
Jensen, A.R. (1985a) 'The nature of the black-white difference on various psychometric tests: Spearman's hypothesis', *The Behavioral and Brain Sciences*, **8, 2,** pp. 193–219.
Jensen, A.R. (1985b) 'The black-white difference in *g:* A phenomenon in search of a theory', *The Behavioral and Brain Sciences*, **8, 2,** pp. 246–63.
Jensen, A.R. (1985c) 'Methodological and statistical techniques for the chronometric study of mental abilities', in Reynolds, C.R. and Willson, V.L. (Eds), *Methodological and Statistical Advances in the Study of Individual Differences*, New York, Plenum, pp. 51–116.
Jensen, A.R. (1986a) 'Individual differences in mental ability', in Glover, J.A. and Ronning, R.R. (Eds), *A History of Educational Psychology*, New York, Plenum.
Jensen, A.R. (1986b) 'The *g* beyond factor analysis', in Plake, B. and Witt, J.C. (Eds), *The Influence of Cognitive Psychology on Testing and Measurement*, Hillsdale, N.J., Erlbaum.
Jensen, A.R. (in press) 'Individual differences in the Hick reaction time paradigm', in Vernon, P.A. (Ed.), *Intelligence and Speed of Information Processing*, Norwood, N.J., Ablex.

Jensen, A.R. and Marisi, D.Q. (1979) 'A note on the heritability of memory span', *Behavior Genetics*, 9, pp. 379–87.

Jensen, A.R. and Osborne, R.T. (1979) 'Forward and backward digit span interaction with race and IQ: A longitudinal developmental comparison', *Indian Journal of Psychology*, **54,** pp. 75–87.

Jensen, A.R. and Reynolds, C.R. (1982) 'Race, social class, and ability patterns on the WISC-R', *Personality and Individual Differences*, **3,** pp. 423–38.

Loehlin, J.C., Lindzey, G. and Spuhler, J.N. (1975) *Race Differences in Intelligence*, San Francisco, Calif., W.H.Freeman.

Paul, S.M. (1984) *Speed of Information Processing: The Semantic Verification Test and General Mental Ability*, unpublished doctoral dissertation, University of California, Berkeley.

Reynolds, C.R. and Brown, R.T. (Eds) (1984) *Perspectives on Bias in Mental Testing*, New York, Plenum.

Roth, E. (1964) 'Die Geschwindigkeit der Verarbeitung von Information und ihr Zusammenhang mit Intelligenz', in *Zeitschrift für Experimentelle und Angewandte Psychologie*, **11,** pp. 616–22.

Samelson, F. (1982) 'H.H.Goddard and the immigrants', *American Psychologist*, **37,** pp. 1291–2.

Scarr, S. and Weinberg, R.A. (1976) 'IQ test performance of black children adopted by white families', *American Psychologist*, **31,** pp. 726–39.

Schmid, J. and Leiman, J.M. (1957) 'The development of hierarchical factor solutions', *Psychometrika*, **22,** pp. 55–61.

Schmidt, F.L., Hunter, J.E. and Pearlman, K. (in press) 'Forty questions about validity generalization and meta-analysis', *Personnel Psychology*.

Scottish Council for Research in Education (1949) *The Trend of Scottish Intelligence*, London, University of London Press.

Sternberg, R.J. (1979a) *Components of Human Intelligence* (Technical Report No. 19), Washington, D.C., Office of Naval Research.

Sternberg, R.J. (1979b) 'A review of "Six Characters in Search of an Author": A play about intelligence tests in the year 2000', in Sternberg, R.J. and Detterman, D.K. (Eds), *Human Intelligence: Perspectives on Its Theory and Measurement*, Norwood, N.J., Ablex.

Vernon, P.A. (1981) 'Level I and Level II: A review', *Educational Psychologist*, **16,** pp. 45–64.

Vernon, P.A. and Jensen, A.R. (1984) 'Individual and group differences in intelligence and speed of information processing', *Personality and Individual Differences*, **5,** pp. 411–23.

Widgor, A.K. and Garner, W.R. (Eds) (1982) *Ability Testing: Uses, Consequences, and Controversies, Part I: Report of the Committee, Part 2: Documentation Section*, Washington, D.C., National Academy Press.