DEVELOPMENTAL REVIEW 9, 234–258 (1989)

REVIEW
Raising IQ without Increasing *g*?

A Review of *The Milwaukee Project: Preventing Mental Retardation in Children at Risk*

ARTHUR R. JENSEN

University of California, Berkeley

The Milwaukee Project: Preventing Mental Retardation in Children at Risk. By HOWARD L. GARBER. WASHINGTON, D.C.: AMERICAN ASSOCIATION ON MENTAL RETARDATION, 1988. PP. xxx + 434.

Infants in the inner city of Milwaukee who were considered at risk for mental retardation because their mothers had IQs of 75 or below were assigned to Experimental (E) and Control (C) groups. From a few months of age to 6 years of age, the E group was given intensive psychological intervention designed to prevent the deceleration in the rate of mental development typically seen in such children. The gains of the E group in Stanford–Binet and Wechsler IQs, as measured against the untreated C group, were considerable, peaking at about 30 IQ points at age 6, when the special intervention ended and the children entered regular school. Thereafter, the E–C IQ difference rapidly decreased, reaching about 10 IQ points by 14 years of age. The E–C difference in IQ was not reflected in the nonsignificant E–C difference in Reading achievement scores or the questionably significant difference in Math achievement scores, on which, by the end of the fourth grade, the mean scores for both the E and C groups were at about the 10th percentile of the normative sample. These results are most plausibly interpreted as a specific training effect of the intervention on the item content of the IQ tests without producing a corresponding change in *g*, the general intelligence factor common to all cognitive tests, that the IQ ordinarily reflects in the untreated population.

The extraordinary experiment that is the subject of this book has been referred to by an obviously skeptical commentator as the "Miracle in Milwaukee" (Page, 1972). A small group of healthy infants born to mothers whose IQs were 75 or below was selected from the largely black "inner city" of Milwaukee for being "at risk" for cultural–familial mental retardation. They were subjected to 6 years of intensive psychological–educational intervention. This unprecedented treatment raised the children's Stanford–Binet IQs (at the age of 6 years) an average of 32 points

Reprint requests should be sent to Arthur R. Jensen, School of Education, University of California, Berkeley, CA 94720.

0273-2297/89 $3.00

above the mean IQ of an untreated control group selected by the same criteria—a difference equivalent to 2.9 within-group standard deviations.

This book by Howard L. Garber, a senior researcher at the University of Wisconsin, is the first (and only) full account of this experiment. Many have waited a very long time for it. Since it was announced as being in the offing for so many years before it was actually published, skeptics had begun to doubt that a bona fide technical report of the study (rather than merely news stories about it) would ever be written and disseminated for critical scrutiny by the scientific community. But now, finally, here it is—presumably the complete "final report" on the famous Milwaukee Project. Most psychologists and educators, along with the general public, have learned about it only through the typically sensational reports in the popular media. The generally unquestioning accounts that have appeared in a number of psychology textbooks have been deplored on the grounds that the technical information provided in the previous nonrefereed publications and in-house progress reports of the study was too inadequate to permit proper critical evaluation or to warrant scientifically worthy conclusions (Sommer & Sommer, 1983).

The complained of information gap is now indeed rectified by the appearance of Garber's impressively detailed monograph, implemented with 69 tables and 43 graphs of summary statistics. The obvious, extreme importance of the personal and social problems addressed by the study seems absolutely unarguable from any humane or rational standpoint: the calamity of mentally retarded mothers bearing and rearing offspring who thereby are deemed "at risk" for mental retardation. The main research question the study was designed to answer is also critically important for psychological therapy and for social action: Can environmental intervention prevent mental subnormality in otherwise healthy children who are seriously at risk for mental retardation or borderline intelligence and poor scholastic performance? Hence we have here the most auspicious conditions for a momentous experiment. And indeed it is. Not too surprisingly, however, the resulting conclusions, as in so many other studies in the behavioral sciences, will be seen as debatable, not so much at the level of the empirical results per se, but at the level of their interpretation.

HISTORY OF THE PROJECT

The project was funded by government agencies for 15 years, from its inception in 1966 to 1981, when the data collection was completed. Since then, Dr. Garber has served as the chief curator of the project. The original director was Dr. Rick Heber, a specialist in mental retardation, on the faculty of the University of Wisconsin in Madison. The project

ended in 1981 at the same time that Director Heber was involved in a scandal, which, although it was not connected with the Milwaukee Project per se, has tended to cast a shadow upon it by association. In striking contrast to the broad national coverage received by the Milwaukee Project in the popular media, the Heber scandal received surprisingly little national coverage by the popular media or in professional publications, although there was quite extensive local coverage of the scandal by a few newspapers in Wisconson. It would be inappropriate here to harp on these surprising misfortunes of Rick Heber and an associate, who had no connection with the Milwaukee Project itself. As I wish to focus on the substantive aspects of the study, I refer readers to the quite detailed history of the project, including an account of the scandal involving Heber, by Ellis Page (1986), a professor of educational psychology at Duke University and a well-recognized methodologist in behavioral research. To date, no outside investigator, to my knowledge, has looked at the substantive and methodological aspects of the Milwaukee Project more thoroughly or critically than Page has over the years (Page, 1972, 1973, 1986; Page & Grandon, 1981). Yet none of Page's articles ever elicited a reply by any of the Project's staff, despite journal editors' invitations. One would have thought that this book would provide the perfect forum for responding specifically to Page's commentaries, but Page is never referred to anywhere in this book. Granted that the rather acerbic style of the questions posed in Page's essays would not be warmly welcomed by the project's staff and supporters, it appears that Page has critiqued carefully, as far as the then limited availability of relevant data would allow.

Whatever else might be said about the Milwaukee Project, it was probably the most expensive single experiment in the entire history of the behavioral sciences—reportedly $14 million (Page, 1986). But it is also a fact that no other attempt had ever before been made at such extraordinarily extensive and intensive manipulation of environmental conditions to promote the mental development of children deemed at risk for mental retardation. Assuming that most of the project's expenditures were intended for this massive intervention, Page estimated that the gain in IQ produced by it cost $23,000 per IQ point per child. Dr. Garber, however, has informed me as follows: "The intervention comprised only a part of a much larger grant, including significant administrative expenditures. The actual monies spent on the stimulation and evaluation of the children over the 20-year history of the project were considerably less than the $14 million quoted by Page (1986)" (personal communication, October 14, 1988).

The scandal involving Rick Heber arose in 1981, when it was discovered that some $165,000 of the federal funds granted to another project

(*not* the Milwaukee Project), of which Heber was also the director, was misappropriated for personal use, resulting in felony convictions for embezzlement and tax evasion, with prison sentences totaling 9 years for Heber and the project's associate director, Dr. Patrick Flanigan. Dr. Heber was fired by the University of Wisconsin, and after serving out his prison sentence and parole did not return to academe. He died in an airplane crash in 1987, while on a safari in Africa.

The implication of the scandal for the scientific integrity and credibility of the study seems a moot issue. People's attitudes will naturally differ on this point, but it seems to me fairest and scientifically most productive if reviewers would examine the study report on its own terms and if their criticism were entirely intrinsic to the material presented. Of course, this examination would include, besides the present book, previous publications and reports that have emanated from the project. Some of these are referenced in Garber's book, but a number of fairly prominent references are missing (e.g., Heber & Garber, 1975, 1980; Garber & Heber, 1977, 1981, 1982), perhaps because the information they contain is more thoroughly covered in the present book.

But the most conspicuous omission is the progress report issued when the subjects of the study were 66 months of age (Heber, Garber, Harrington, Hoffman, & Falender, 1972). It is by far the largest (247 pages) and most informative report prior to the present book, and it contains test data and valuable details about the training procedures and conduct of the project that were not included in the present volume. One of the repeatedly used tests (Cattell Test of Infant Intelligence) in the 1972 report was dropped from the 1988 book without explanation. The actually trivial discrepancies between similar figures in the 1972 and 1988 reports are attributable to the fact that in the 1972 report the Cattell IQs were averaged in with the Stanford–Binet IQs, whereas in the present report (1988) only the Stanford–Binet IQs are presented. But it seems puzzling that any psychometric data worth reporting in 1972 should have been discarded in the final report. Mental measurements derived from different standardized tests given concurrently, it seems, should aid our understanding of the results of the experiment.

THE NATURE/NURTURE ISSUE

Garber wisely soft pedals the nature/nurture issue, as did Heber before him. But there are scant explicit disclaimers related to this matter and so some readers who are less sophisticated about behavioral genetics than Garber (and Heber) are apt to draw unwarranted inferences from the principal findings of this study. In fact, the Foreword by J. McVicker

Hunt places so much emphasis on what he apparently supposes are the study's implications for the passé nature/nurture controversy that it seems necessary to set things straight on this score before going on to consider the true gist of the study.

Hunt mistakenly believes that the findings of the Milwaukee Project constitute evidence against, in his words, "Jensen's contention that children of mothers with phenotypic IQs of less than 80 are doomed to have IQs under 80" (p. xxiv). But I have never made any such claim. In the first place, such a claim would be obviously incorrect in any case, just in terms of basic principles of quantitative genetics. If the mothers' IQs were below the average of their population, their offspring's IQs would regress toward the population mean and consequently would have a somewhat *higher* average value than that of their mothers. In the second place, neither the observed (or phenotypic) parent–offspring correlation for IQ nor the theoretical genetic parent–offspring correlation would permit anything other than a statistical, or probabilistic, prediction of any given offspring's IQ from a knowledge of its parent's IQ. Hence any particular child is not "doomed" to an IQ below 80.

In fact, there is absolutely no evidence presented in Garber's monograph that either contradicts or is inconsistent with our present knowledge of behavioral genetics or the heritability of IQ. The experimental design itself only permits a test of the hypothesis that the effect of the environmental intervention either is or is not significant at some specified level of confidence, as measured against an untreated control group. If the magnitude of the effect turned out to be significantly larger than would be predicted from a well-established estimate of the heritability of IQ (on the basis of twins, a variety of other genetic kinship correlations, and adoption studies), it would mean that certain environmental influences had been brought to bear that exceeded the range of environmental effects in the general population in which the heritability of IQ had been estimated. This would be an important discovery, assuming that these exceptional influences can be described and replicated.

But how can we estimate the range of environmental effects on IQ from the best statistical estimates of its heritability in the general population? Heritability, a technical term symbolized as h^2 in quantitative genetics, is the proportion of population variance in a given trait attributable to genetic factors. The proportion of nongenetic variance, then, is $1 - h^2$. If the population standard deviation (*SD*) of IQ is 15, the variance is $(15)^2 = 225$. The overall average of the best estimates we have for the broad heritability of IQ is between .60 and .70. For this example, let's assume the higher figure. Then the proportion of the total IQ variance attributable to nongenetic effects (i.e., environment and measurement error) would be

1 − .70 = .30. The distribution of these nongenetic IQ deviations from the population mean, then, would have a *SD* of $\sqrt{.30\ (225)}$ = 8.21 IQ points. Research on monozygotic twins reared apart and reared together indicates that the nongenetic effects on IQ have an approximately normal distribution. Given the above conditions, then, we can estimate what geneticists refer to as the *reaction range* of IQ, that is, the range of variation in IQ in the natural environment if all genetic variation were removed. The reaction range of IQ, assuming a broad heritability of .70, can be defined as the total range of IQ between the 1st and the 99th percentile in the normal distribution of environmental effects in the general population. Given the stated conditions, this range amounts to approximately 38 IQ points. This means that on a continuum of the total effect of environmental influences on IQ, the most favorable 1% results in an IQ that, on average, is 38 points above the least favorable 1%. Hence, environmental interventions that produce effects which are not significantly greater than 38 IQ points are consistent with a heritability of .70 for IQ, without positing environmental factors different from those that contribute to the IQ variance in the normal range of environmental variation in the population. If we can believe that the environmental conditions for the development of IQ that separate the experimental and control groups during the 6 years of this experiment are as different as the most favorable 1% and the least favorable 1% of environments in the general population with respect to the development of intelligence, the outcome of the experiment is not at all inconsistent with present evidence on the heritability of IQ.

Nevertheless, IQ differences of the magnitudes found between the Experimental (E) and Control (C) groups in the present study are still extraordinary, assuming that the experimental treatment has produced the large E–C difference in IQ without altering the essential nature and meaning of the IQ itself. The main reason it is so extraordinary rests on our present view of the nature of the substantial environmental variance in IQ, which constitutes some 30 to 40% of the total IQ variance.

In the past decade, evidence from behavior–genetic analyses of IQ, mainly from adoption studies, has led some behavioral geneticists (e.g., Plomin & Daniels, 1987) to argue that the major part of the nongenetic variance in IQ, especially after puberty, is attributable to *nonshared* environmental influences within the family. The *shared* environment (also termed the *between-families* environment or *common* environment) comprises all the nongenetic influences shared by all the children in a given family but which differ *between* different families in the population. The nonshared environment (also termed *within-family* environment or *specific* environment) comprises the nongenetic influences that differ between children reared together in the same family. Examples of the

shared environmental effects on children's mental development are their parents' intelligence, education, socioeconomic status, culture, dietary habits, life-style, and the like. Examples of the *nonshared* environment are differences between siblings (or unrelated children reared together in the same family) in birth rank, idiosyncratic prenatal and perinatal conditions, different treatment of siblings by the parents or others, different health histories, and different experiences outside the home.

One of the striking pieces of evidence that shared environment has little effect on IQ is a study by Scarr and Weinberg (1978) of a large sample of unrelated pairs of adolescents who had been adopted before 1 year of age and reared together in families ranging widely in socioeconomic levels from blue collar to professional and managerial. Since the adopted pairs were genetically unrelated, any resemblance between them in IQ had to be due to selective placement by the adoption agencies and to shared environmental influences. But the correlation between their Wechsler IQs at about age 18 years was only a nonsignificant $r = -.03$! (The IQ correlation between preadolescent adoptees reared together is close to $r = +.20$.) In other words, by late adolescence the effect of shared environment (or of environmental differences between families) was absolutely nil. Scarr and Weinberg (1978) concluded. ". . . intellectual differences among children at the end of the child-rearing period have little to do with environmental differences among families that range from solid working class to upper middle class" (p. 691). The substantial variance in the adoptees' IQs attributable to environmental factors consisted entirely of nonshared environmental effects.

This is an important point, because one of the unresolved questions in this field concerns the nature of the nonshared environment. The variance attributable to the nonshared environment has also been labeled "unsystematic" and "microenvironmental," implying that this component of the IQ variance reflects mainly the effects of a great many very small, largely uncorrelated physical and psychological influences. In theory, the various lowly correlated influences could each be so small and yet so numerous that their combined effects would practically constitute a normally distributed random variable, in accord with the Central Limit Theorem.

Now if that model of the nonshared environment is true, it should be exceedingly difficult, if not practically impossible, to purposefully manipulate the nonshared environment so as to produce a large positive increment in IQ. The total environment would have to be so controlled that what would ordinarily be a great many small and scarcely correlated effects would all be caused to act together consistently in one direction favorable to intellectual development. Of course, because of the laws of chance this would actually happen to some very small proportion of in-

dividuals in the population—those who have "lucked out" with respect to the nonshared "microenvironment." With such luck, an individual could have an IQ perhaps 20 to 30 points higher than if he or she had equally bad luck with respect to the nonshared environment.

Hence, to the extent that the nonshared environment ordinarily acts like a random variable in its effect on IQ, it would imply inordinate difficulty for the intentional improvement of the mean level of IQ in a *group* of individuals by environmental means—*unless* such means included unusual influences that normally do not contribute appreciably to the environmental component of the IQ variance. That would be one reasonable interpretation of a treatment that markedly improved IQ, assuming that the raised IQ retained the same meaning (i.e., construct validity) it would have in an untreated but otherwise comparable group.

Another reasonable interpretation is that the nonshared environmental variance is really not attributable to such "microenvironmental" and "unsystematic" influences as to be virtually a random variable, but is mainly the result of just a few diverse but highly important influences that are amenable to intentional control. Since these few crucial influences on intellectual development have not yet been specifically identified, they have not all been fully brought to bear in the countless previous attempts to raise intelligence. An optimal combination of such crucial influences perhaps could be the "active ingredient" among the incredibly immense variety of cognitive experiences, training, and adult–child interactions provided to the experimental group in the Milwaukee study.

Still another plausible hypothesis is that the shared family environments of the children selected for this study were generally so extremely unfavorable for intellectual development as to be almost completely outside the range of environments in the general population in which the existing analyses of genetic and environmental components of IQ variance have been done. In that case, there is little basis for predicting the magnitude of the effects that any particular improvements in the environment will produce. And the home environments of the subjects in the Milwaukee Project certainly appear exceedingly impoverished with respect to the conditions that psychologists generally believe foster mental development.

A final caveat about heritability. In quantitative genetics, the available methods for the estimation of heritability are all essentially forms of the analysis of variance. Hence, the results of any heritability analysis are necessarily limited to statements concerning variation around the overall *mean* of the group in which the analysis is performed, and it affords no information whatsoever about the factors responsible for the particular value of the group mean. Adding a numerical constant of any size to all of the individual measurements that are entered into a heritability analysis,

although it would change the overall group mean, would have no effect whatsoever on the estimated proportions of genetic and nongenetic variance estimated by the analysis. Thus, the concept of heritability, as technically defined and determined in quantitative genetics, is simply uninformative concerning the casual nature/nurture aspect of the sample (or population) mean. For that reason, the methods of quantitative genetics cannot tell us anything about the relative contributions of genetic and environmental factors to the overall mean IQ of the mothers whose children were the subjects of this experiment.

EXPERIMENTAL DESIGN

Garber devotes 45 pages to describing the subject selection, design, and nature of the experimental treatment in this longitudinal study. Without wishing to minimize the importance of the investigators' rationale for all these aspects or the details of the procedures (sample attrition, and the like) and practical logistics that led to the final conditions of the experiment, I can indicate here only the features that are most essential for conveying the gist of the study.

Subjects. Selection of all of the subjects in the study was limited to black families in the poorest "inner city" area of Milwaukee. Mothers with newborn infants were screened on Wechsler Vocabulary, and if they scored more than one *SD* below the mean scaled score they were given the entire Wechsler Adult Intelligence Scale (WAIS). If their Full Scale IQ was 75 or below, they were asked to volunteer participation of their infants in the project. Infants selected on this basis would ordinarily be expected to show a decline in cognitive ability during the preschool years, have problems in school, and be at risk for eventual identification as cultural–familial mentally retarded. Selection was limited to only physically normal, healthy infants.

A total of 40 such infants was finally recruited; 20 were assigned to the experimental (E) group and 20 to the control (C) group. A few families later dropped out of the project, leaving E = 17 and C = 18 that remained throughout the study and on which virtually all of the main data summaries were based. Because the task of finding qualified subjects who were likely to remain in the area for the duration of the project necessarily extended over a period of 24 months (1966–1968), plus the importance of beginning the full treatment for the E subjects as early as possible and never later than 6 months of age, true random assignment to the E and C groups was simply not feasible. So subjects were assigned to the E and C conditions on an alternating basis as they were recruited. I see no real problem with this procedure. Comparisons of the E and C families on various items of information on the mothers (age, IQ, education, reading

level, number of children) and on their infants (birth rank, spacing, length, and weight) indicate that the E and C families are reasonably comparable for the purposes of this study. A few statistical perfectionists might quibble, but a counsel of perfection hardly seems reasonable for a study of such unusual practical difficulties that it is even a wonder it could have been done at all.

Also obtained from the same area was a *contrast* group of infants (N = 8) considered to be at "low risk" for mental retardation, since they were selected for having mothers with WAIS IQs of 100 or above.

When the mothers whose children completed the program were initially tested for selection into the study, their mean Full Scale IQs were E group = 68.12 (SD = 6.37), C group = 65.94 (SD = 6.99), and Low Risk (LR) group = 111.29 (SD = 3.99). Their years in school averaged approximately 10, 9 and 13, respectively, for E, C, and LR. A predictable regression effect on the mothers' IQs is seen on retest 6 years later, when the study was completed. That is, the IQs of the E and C group mothers, who had been initially selected for low IQ, were slightly higher and the Low Risk mothers (who were initially selected for high IQ) were considerably lower than on the initial test, as follows: E = 72.62 (SD = 7.05, C = 71.22 (SD = 7.54, and LR = 100.29 (SD = 4.99). The consistent increase in the SDs from test to retest also reflects a predictable regression effect. Hence the mothers' retest IQs probably represent the more accurate estimate of their intellectual status in terms of the Wechsler norms.

The mothers of the E and C groups came from families averaging 9.3 children. By the conclusion of the study, they themselves averaged 5.1 offspring. (In striking contrast, the 1970 U.S. Census shows that black women who have attended college for 3 or more years have an average of 1.3 offspring by the age of 44.)

Experimental intervention. This had two components: (1) *family/maternal rehabilitation* provided by project staff visiting the homes of the E group for 3 to 5 h 3 days a week during the first 18 months of the program, for instructing the mothers in child care, nutrition, hygiene, money management, social services, and the like, in addition to half a year of remedial education and vocational training for the mothers; and (2) continuous *infant/early childhood stimulation* until the children were 6 years of age and entered the first grade in public school. At some variable time before they were 6 months of age, the E infants were removed from their homes for a large portion of every day, to be spent in the Infant Stimulation Center that was established in the heart of the "high risk" neighborhood.

The key hypothesis of the study was that a retarded mother creates a "psychosocial and intellectual microenvironment" that is inadequate for

stimulating a normal rate of cognitive development in her offspring. Hence, in the Stimulation Center, specially trained paraprofessionals were employed to act as surrogate caregivers, who assumed responsibility for the total daytime care of the infants. A core group of nine caregivers worked in the project for more than 5 years. During the infancy period the adult:child ratio was 1:1; later the ratio gradually changed to 2:2 and then to 1:3.

The kinds of physical and mental stimulation and training given to the E children during their 6 years in the Stimulation Center are far too varied and numerous to be listed here. They include more ways that might conceivably promote cognitive development than a team of child psychologists could think up given a month with nothing else to do. Just about every form of didactic stimulation ever suggested by child development experts from Montessori and Piaget to J. Mc V. Hunt seems to have been scheduled. In 1967, when I first met Rick Heber and was introduced to the Milwaukee Project, I was told that a large part of his research staff's activity at that time was devoted to generating methods and materials for promoting cognitive development. It takes 24 book pages just to list all of the main features of the stimulation/training program.

The impression I get from the description of this program is that much of the training activity reflects a behavioristic conception of cognitive development as consisting of the acquisition of the particular kinds of knowledge and skills that psychologists would glean from a task analysis of the various items typically found in standard intelligence tests. This is not a bad basis for training cognitive abilities. What else would one have thought of prescribing for what was intended to be an unprecedented comprehensive regimen for maximizing intellectual development? (I myself cannot think of any activity that might have been helpful that was not included.)

But such all-inclusive and intensive training does raise the critical question, in this type of experiment, of "training to the test," however unintentional or inadvertent. It is surprising that this problem did not seem to have been a consideration, much less a worry, to the principal investigators. If there was any explicit recognition of it in any of the project publications, I can't recall it. Admittedly, it would take some ingenuity, even more so in the 1960s than today, to figure out how to avoid "teaching to the tests" while providing an exceptionally comprehensive program for cognitive development, with prior knowledge that the effects would be assessed periodically with such familiar standardized tests as the Gesell Developmental Schedules, the Stanford–Binet, and the Wechsler (WPPSI and WISC).

Although subjects were not trained on actual test items, of course, it appears from the specific elements in the training program that it would be

practically unavoidable that a good many of the test items would likely fall within the rather narrow generalization or transfer gradients of the many specific skills that had been assiduously taught and practiced. There is nothing wrong with that in itself, of course. But from a theoretical standpoint the issue is important for interpreting the nature of the treatment effect. If we accept a behavioristic definition of intelligence as the repertoire of specific bits of knowledge and skills in the domain from which the items in the Stanford–Binet and Wechsler IQ tests are sampled, there is little problem. But that specificity theory of intelligence has been largely abandoned by cognitive theorists in the last decade. Under the circumstances, the effect of the cognitive stimulation program would be best assessed, in the final analysis, in terms of its practical consequences for the treated subjects in meeting the "real life" demands that normally reflect general cognitive ability, such as scholastic performance.

To suggest that the stimulation program these children were exposed to is similar to the microenvironment that middle class parents or mothers with IQs >100 typically provide for their children would seem a gross understatement. I recall a conference at which Heber expressed his belief that probably no children previously had ever spent so much time in such an intensely stimulating environment for intellectual development as did the Milwaukee Project's E group. He quipped that the childhood environments of John Stuart Mill and Sir Francis Galton would seem very deprived by comparison.

The daily program of the Stimulation Center continued until the E children were 6 years of age and entered the first grade in ordinary public schools. To prevent their being surrounded by low-achieving classmates, efforts were made to place them in schools that had fourth grade achievement test scores near the national and citywide averages.

The Control and Low Risk groups, of course, received none of these services. All three groups, however, were periodically given the same tests at the same ages. Also, one IQ test was obtained on most of the subject's siblings, thus permitting other informative comparisons.

PSYCHOMETRIC ASSESSMENTS OF TREATMENT EFFECT

During the 6-year intervention period, the following standardized tests were individually administered to each of the subjects in the E, C, and Low Risk groups: Gesell Developmental Schedules, ages 10 to 22 months (four times); Stanford–Binet, ages 24 to 48 months (nine times) and 54 to 72 months (four times); Wechsler Preschool and Primary Scale of Intelligence (WPPSI), ages 48 to 72 months (three times). Because of the repeated administration of the same tests (e.g., 13 times for the Stanford–Binet) and the unassessed magnitude of practice effects, one cannot put a

great deal of stock in the absolute values of the IQs for any of the groups. As Garber notes, one can really rely only on comparisons between groups. These are meaningful, as all groups received all the same tests on the same age schedule.

Most of the tests were given by one and the same examiner on the project's staff who was not involved in the educational program. Garber states, "Examiner bias in Stanford–Binet IQ scores was assessed by comparing scores obtained by the project examiner from the 48- through 72-month assessments with scores obtained during the same general period by independent examiners brought in on three separate occasions during the same period for this purpose" (p. 91). Although the IQs were highly correlated ($r = .96$) across testers, the IQs obtained by the project tester averaged 7 points higher for the E subjects and about 5 points higher for the C subjects, but this has little effect on the mean E–C difference of about 28 IQ points.

The information about the testing procedures, however, seems rather sketchy. It is unclear how much of the testing was done "blind" with respect to subjects' E or C identity (see p. 76); no information is given about the exact number of independent examiners, or just how many of the E and C subjects they tested, or the *N*s of the correlation and means reported to show the comparability of the project examiner and the independent examiners. All that I can find on this score in the earlier report by Heber et al. (1972) states,

> . . . to evaluate the effects of the examiner in the production of group differences in test performance, one test administration was conducted by a qualified examiner brought from a neighboring state for this purpose. All subjects were tested in an environment totally unfamiliar to them and the examiner was not apprised of the subject's group membership. At that time, all four-year scores were complete and comparison of these scores with the independent tester showed no significant difference. (p. 50, including a table of means and *SD*s)

The main point. There were significant mean E–C differences in the age-standardized scores (DQ or IQ) on all of the tests given in the first 6 years. Scores of the Low Risk contrast group were generally lower than those the E group but were well above those of the C group; but the Low Risk group, with ($N = 8$, is statistically far from ideal and is obviously at high risk for Type II Error (i.e., not rejecting the null hypothesis when it is false). Nearly all of the E–C divergence in scores occurs on the Gesell Developmental Quotient (DQ) between 10 and 22 months of age—reaching a difference of about 25 DQ points. This large E–C difference appears to be nearly stabilized at about 2 years of age, and the experimental treatment maintains approximately this same relative difference on the Stanford–Binet IQ rather consistently up to 6 years of age. The

Stanford–Binet IQs show a fairly constant E–C difference (averaging close to 30 points) across the 13 testings from 24 to 72 months of age.

The many statistical analyses to which these data were subjected leave no doubt that the E–C mean differences in DQ and IQ are highly significant at every age level. To give readers a general overview of the IQs in the various groups in the whole study, I have used all of the IQ data presented (in Chap. 11) on the individual subjects and their siblings to calculate the statistics shown in Table 1. "Target" refers to the study samples that were periodically tested, and so their IQs therefore may exceed their siblings' IQs to some unknown degree because of the practice effect. The IQs in Table 1 were obtained in 1974, when the study subjects were between 6 and 8 years of age; the sibling group was, of course, more variable in age. The mean 5.13 IQ points difference (significant at $p < .05$) between the E and C siblings is presumably attributable to the effects of the family rehabilitation program on the E group mothers. Although Garber notes that the initial superiority of the final E group mothers in education, IQ, and literacy levels (Table 3–2, p. 35) is not statistically significant, it would have been more informative to have shown what proportion of the variance between the E and C siblings' IQs is accounted for by these maternal characteristics when they are entered into a multiple regression.

Learning and language assessments. Besides the standardized tests, the E and C groups were compared at yearly intervals between 30 and 102 months of age on each of three learning task paradigms: color–form matching, oddity discrimination, and probability matching. There were fully significant differences on the first two and a mixture of significant and nonsignificant differences on variables derived from the third task, which is beset by problems of analysis. Unfortunately, comparisons between the E–C effect size (in terms of within-group *SD* units) on the standardized tests and on the learning tests would not be legitimate (and hence were not made) because of what appears to be a ceiling effect on the scores of the E group, making it indeterminable whether the educational intervention caused the E and C groups to differ more (or less) on

TABLE 1
IQ STATISTICS ON GROUPS IN THE MILWAUKEE PROJECT

	Experimental		Control		Low risk	
	Target	Sibling	Target	Sibling	Target	Sibling
N	21	74	17	52	8	9
Mean	107.5	85.0	81.2	79.9	105.9	96.7
SD	10.1	10.2	8.7	15.1	11.8	10.9

these learning tasks than they differ on the standard IQ tests—a point with theoretical implications for interpreting the nature of the intervention on IQ.

The intervention program placed great emphasis on language development, and periodic E–C comparisons were also made on a wide variety of language assessments. E was almost invariably superior on these measures. Probably the most comprehensive measure in this domain is the Illinois Test of Psycholinguistic Abilities, which resembles the Stanford–Binet IQ in the overall size of the E–C difference.

Mother–Child interactions. One chapter is devoted to assessments of E–C differences based on the many systematic observations of mother–child interactions in a variety of controlled situations. The conclusions stand most earlier views of mother–child interaction on their heads. Here it is emphasized that the characteristics of the child itself, more than the mother, control the quality of the interaction, particularly in verbal communication and teaching situations. Garber observes, "Although the direction of the interaction seemed mainly to be from the mother to her child, it was actually the child who was manipulating the interaction" (p. 201). E and C mothers were virtually forced to interact differently with their children in certain situations because of differences between the behaviors of the E and C children. These observations reinforce the growing realization by developmental psychologists that the child's social environment is at least as much a reaction to the child's own characteristics as it is a molder of them (Scarr, 1985; Scarr & McCartney, 1983).

Postintervention follow-up assessment of IQ. The last in the longitudinal series of 13 Stanford–Binet IQs was obtained at age 6 years. At ages 4, 5, and 6 years, WPPSI IQs were also obtained. Because of repeated testing and unknown practice effects, the E–C comparisons are more meaningful than the absolute values of the IQs. Practice effects for the E and C groups are presumed equal.

There are some telling features of these data. The Stanford–Binet (S–B) IQs at age 6 were E = 119, C = 87, mean difference = 32. At the same age (6 years), WPPSI IQs were E = 109, C = 88, mean difference = 21. Note the large (10 points) difference between the S–B and WPPSI IQs for the E group and the negligible 1 point difference in the C group. Also, WPPSI IQ shows a decline of 4.94 points (more than half a *SD*) from ages 4 to 6 in the E group as compared with a negligible gain of 0.25 point in the C group. The exclusively E group difference between the S–B and WPPSI at age 6 and the E group's decline in WPPSI IQ between ages 4 and 6, viewed together, suggest the hypothesis that intervention produced greater specific transfer effects to the S–B than to the WPSSI, but this hypothesis cannot be tested fairly with such a small *N*.

The Wechsler Intelligence Scale for Children (WISC) was obtained at

ages 7, 8, 9, 10, 12, and 14 years. At age 7, Groups E and C had mean Full Scale IQs of 103 and 81, respectively—a difference of 22 points. At age 10, the mean IQs for E and C were 104 and 86—a difference of 18 points. And there continues a gradual decrease in the E–C difference. By the last testing, at age 14, the E and C Full Scale IQs were 101 and 91—a difference of 10 points. (The Low Risk Contrast group, with $N = 5$, had an IQ of 97 at age 14.) In Chapter 11, the Stanford–Binet and WISC IQs on individual subjects are shown graphically at every test administration. In scanning these graphs, one notices that in the E group the WISC IQs (obtained after intervention) are quite consistently lower than the S–B IQs (obtained during intervention), whereas in the C group the S–B and WISC IQs are generally about the same. (I tested the E–C difference in this respect by means of χ^2 which is significant at $p < .01$.) Again, the transfer effect of intervention on the IQ is significantly reduced when assessments shift from the Stanford–Binet to the WISC.

If we look at the mean E–C differences in units of the average within-group *SD*, the WISC Full Scale IQ differences at ages 7, 10, and 14 years are 2.24 *SD*, 1.64 *SD*, and 0.87 *SD*, respectively. The largest E–C mean IQ difference appeared at age 6, at the end of the intervention program, when the Stanford–Binet IQs for E and C were 119 and 87—a difference of 32 points, or 2.92 *SD* units. Still, a 10-point IQ difference (equivalent to 0.87 *SD*) remained at age 14, a full 8 years after the conclusion of the intervention. Moreover, at no age did the E group have any members with IQs of 80 or below, while at age 12 and above, 41% of the C group and 35% of the untreated siblings of the E group had IQs of 80 and below. This is indeed remarkable, provided we can assume that the E–C difference in IQ actually reflects the same degree of difference in the latent ability factor (for which the IQ is merely an index) as that represented by a comparable IQ difference between untreated groups in the general population.

We see an interesting feature when we look at the WISC Verbal and Performance IQs separately. In the national standardization data, the Verbal and Performance IQs differ less than 1 IQ point, for blacks as well as for whites (Jensen & Reynolds, 1982). But in the present samples, especially the E group, from age 7 on, Performance IQ exceeds Verbal IQ, a trend that by ages 12 and 14 results in conspicuous differences. At age 14, for example, we see the following:

	Verbal IQ	Performance IQ	*P–V*
Group E	93	109	16
Group C	85	99	14
E–C	8	10	

What can it mean, that the Performance IQ–Verbal IQ differences *within* each group are even larger than the differences *between* the E and C groups? The within-group difference suggests a practice effect due to repeated testing, which is much greater on the Performance tests than on the Verbal tests. This is consistent with expectations, since some of the Performance tests, such as Block Design and Object Assembly, remain essentially the same at every administration and rises in score can be obtained simply by an increased speed of performance, which is facilitated by familiarity and practice. In the case of the Verbal subtests, such as Information and Vocabulary, however, new items are presented on successive administrations, assuming there is some growth in vocabulary during the intervals between tests. The specificity of Information and Vocabulary knowledge, once learned, can have little, if any, transfer or practice effect to other items, whereas practice on one Block Design problem transfers to subsequent Block Design problems, and the same is true for other Performance subtests, such as Object Assembly and Digit Symbol (or Coding).

This observation is important in the present context because it indicates that various tests differ in sensitivity to training, practice, or transfer effects. And if such effects can produce such substantial differences between subtests *within* a group, it is reasonable to suppose that similar effects of the intervention treatment could account for a substantial part of the test score difference *between* the treated (E) and untreated (C) groups.

The crucial question, then, is do such effects change the meaning of the scores on the tests on which the effects occur? We know that in the Wechsler standardization population the Verbal and Performance IQs are both very highly loaded on the general factor, or *g*, which all of the subtests have in common (e.g., Jensen & Reynolds, 1982). The *g* factor emerges as the highest order common factor in a hierarchical factor analysis of any large and diverse battery of cognitive tasks (Jensen, 1987a). The high *g* loadings of both the Verbal and Performance scales is reflected in a correlation of close to +.80 between the Verbal and Performance IQs in the standardization population. Thus the expected value of the Verbal IQ as predicted from the Performance IQ (or vice versa), on average, should differ only slightly from that of the Performance IQ.

Now, when a group's mean is relatively inflated on the Performance IQ as compared to the Verbal IQ (or vice versa), two questions call for answers: (1) Do both scales measure the same *g* factor to the same degree as in the general population? and (2) Does the level of IQ on the relatively more inflated scale have the same meaning with respect to external criteria as the same IQ score in the general population? Since in the WISC

standardization population both the Verbal and Performance scales reflect *g* more than they reflect any other sources of variance measured by the various Wechsler subtests (including the Verbal and Performance factors independent of *g*), a large discrepancy between the mean Verbal and Performance IQs for a given group (when no such discrepancy exists in the population) raises the question of whether the higher Performance IQ actually represents a higher level of *g* or merely a higher level on non-*g* sources of variance that are entirely specific to particular Performance subtests (technically termed the *specificity* of a test). If the latter is the case, then, of course, whatever caused the inflated score will result in little or no transfer or generalization to other *g*-loaded criteria that do not share the same specificity.

The crucial distinction between psychometric g and its vehicle. A critic, of course, always has the advantage of hindsight, and what I am emphasizing now would probably not have occurred to me in the Zeitgeist of 1965–1966, when the Milwaukee Project was in the planning stage, any more than it occurred to the original investigators, apparently. But it now seems evident that the study invested entirely too much stock in too few and too similar tests, mainly the Stanford–Binet and Wechsler scales, although both are known to be highly *g*-loaded in the general population. Data from a greater variety of *g*-loaded tests that have little superficial resemblance to one another or to much of what was specifically trained in the intervention would have been much more informative as to whether the intervention increased *g* or increased only the specificity of the particular vehicles used to measure IQ.

It is entirely possible to raise IQ scores on specific tests without raising *g*. Test items are merely vehicles for the measurement of *g*. Although the *g* factor ordinarily constitutes the largest proportion of the variance of the total scores on IQ tests, *g* accounts for only a small fraction of the total variance on any given item, and it is possible to train up the specificity of items or of particular subtests composed of similar items without having any effect on *g* itself. The resulting increment in the total test score, then, does not reflect the individual's true level of *g*. The added increment is merely a non-*g* inflation of the test score. Such a test score reveals its inflated status by *over*predicting the level of performance on some independent *g*-loaded criterion (or test). That is, the person whose test score is raised by inflation with non-*g* factors will not perform up to the same level on other criteria as another person who has obtained the same score without such inflation.

This phenomenon is seen most dramatically perhaps in studies in which digit span memory has been specifically practiced over a period of time. Ordinarily, unpracticed individuals who score, say, 1 *SD* above the pop-

ulation average on *digit* span also score nearly 1 *SD* above the average on *letter* span. But if a person who is just average when first tested on digit span then practices on digit span tests for a time, the person's digit span score can be raised to 2 or 3 or more *SDs* above the average, depending on the amount of practice. But then, when this person is tested on letter span, the performance is found to be just about average (Ericsson, 1988). The specificity of digit span memory, rather than the more general ability it normally reflects (in addition to the specificity), is all that had been improved by practice. The same thing can happen for practically every specific vehicle used to measure *g*, or general intelligence, which includes all of the conventional IQ tests. But if the present study had employed a wide variety of dissimilar vehicles that have known *g* loadings in some untreated population, we would be in a much better position to discover the degree to which the intervention affected *g*, and not just the specificity of the particular vehicle used to measure it. Today I would suggest using a variety of reaction time and inspection time techniques that are known to reflect *g* while involving very little or no intellectual content. And these measures of the speed of information processing bear virtually no resemblance to the usual psychometric tests of intelligence. The two classes of tests have practically no factors in common other than *g*. Various reaction time measures, for example, show highly significant differences between groups that clearly differ in IQ (Cohn, Carlson, & Jensen, 1985; Jensen, 1982, 1985a, b; Jensen, Cohn, & Cohn, 1989). How much of a difference would they have shown between the E and C groups of the Milwaukee Project? If the difference were comparable to that found between two untreated groups with the same mean IQs as the E and C groups, we could be reasonably confident that the level of *g* itself, and not just Stanford–Binet or WISC IQ scores, had been raised by the intervention.

We should not belittle the practical value of training up specific skills, however. That is a large part of what education is about, and any one of us would be bad off indeed were it not for all of our trained and highly practiced skills. But while such acquired knowledge and skills are valuable in their own right, they do not have the essential properties of *g*, which is the sine qua non of IQ and is the chief "active ingredient" responsible for the practical predictive validity of IQ tests. With *g* partialled out, the residual IQ would measure virtually none of the extremely generalizable mental ability that has made the IQ so important. The IQ became controversial mainly because it purports to measure one of the most highly valued human traits. The variety of the nonpsychometric correlates of IQ, which depends largely on the extent to which the IQ validity reflects *g*, comes as a surprise even to many psychologists (Jensen, 1984, 1985a, b, 1986, 1987a, b). Indeed, IQ is really important

only to the extent that it reflects g. And g is important because of its correlation (and causal connection) with a host of educational, occupational, economic, and social variables that are highly valued in every industrialized society in the world.

SCHOLASTIC ACHIEVEMENT OF THE E AND C GROUPS

Scholastic achievement, as validly measured by objective tests, is known to be highly g loaded and is quite highly predictable by IQ. Also, the school curriculum at each successive grade level departs further from the specific abilities that were cultivated in the E group during the intervention program. Therefore, a comparison of E and C groups in scholastic achievement is probably the best evidence this study affords of the intervention effect on g and whether the statistically significant intervention increment in IQ (as measured by E–C) is or is not "hollow" with respect to g. We already know from numerous studies that untreated groups in the population that differ in mean IQ also differ to a very nearly equivalent degree in mean level of scholastic achievement. The black and white school populations, for example, differ by approximately 1 *SD* in IQ and they also differ about 1 *SD* in scholastic achievement. The size of the black–white achievement difference reflects the degree of g loading of the particular school subjects tested—hence the difference is larger in reading comprehension and arithmetic problem solving than in spelling and arithmetic computation.

The C group entered regular kindergarten at the usual age (about 5.5 years), while the E group remained in the Stimulation Center for another year, where they were given reading instruction beyond that usually offered in the public school kindergarten. The E group also attended a special summer school tutorial program for 6 weeks, which emphasized instruction in reading and arithmetic, just before entering the first grade.

Both the E and C groups entered first grade in the public schools of Milwaukee at the same age. But only the E children received whatever benefit might be obtained from placement in schools known for a relatively good level of achievement, while the C children attended the nearest neighborhood school in their poor section of the city. At the time of school entry it became obvious to teachers (and also confirmed to them by a school psychologist) that the E children had had previous schooling. Presumably, the C children were not singled out by their teachers.

The Metropolitan Readiness Battery, administered before the beginning of first grade, showed the E group clearly ahead of the C group, with mean readiness percentiles of about 76 and 30, respectively.

The Metropolitan Achievement Tests (MAT) were routinely administered by the schools toward the end of each school year, and the scores

were available on the E and C groups for grades 1 through 4. The MAT covers Reading (word knowledge, word analysis, reading, spelling) and Math (computation, concepts, problem solving). The E and C scores at each grade are reported by Garber in the form of percentiles and grade equivalents. But no *SD*s are given, and so we cannot compare the E–C differences on the MAT with the WISC IQ differences (obtained at the same ages) in terms of within-group *SD* units.[1]

Reading achievement is probably the most telling, because it is usually the most highly correlated with IQ, the most *g* loaded of the school subjects, and also the most predictive of later academic performance, being the chief medium of advanced learning. The MAT Total Reading percentile for the E group rapidly declined from 48.71 at the end of first grade to 19.00 at the end of fourth grade. The corresponding scores for the C group were 31.53 and 8.82. The observed Reading scores are higher for E than for C, but a MANOVA test of the overall differences between groups E and C in grades 1 through 4 shows the difference to be statistically nonsignificant (the *F* and *p* values are not reported). Some readers may be tempted to indulge in the common statistical fallacy of believing that if the sample sizes had only been a good deal larger, the observed differences would be significant. But what statistical nonsignificance means, of course, is that if the *N* were considerably increased, the observed nonsignificant difference could just as well as not vanish altogether. Every seasoned researcher has had this usually disappointing experience. Garber summarized the results as follows:

> . . . they [the E group] did not make progress as fast as did the children who comprised the norm group used to standardize the MAT Batteries. The deterioration in percentile ranks is quite marked and was unanticipated, especially for the experimental group, which had a mean readiness percentile of 75.94 on the MRT [Metropolitan Readiness Test] just before entering school . . . the deterioration in performance began the first year of school and was probably more severe for experimental than for control children." (p.264)

The MAT Total Math percentile for group E declined from 33.51 in first grade to 10.63 in fourth grade; the corresponding C group scores

[1] Since writing this, Dr. Garber has provided me with a table containing both the means *and* the *SD*s of the MAT percentiles for Reading and Math. Because the variances of the C group are only about one-third those of the E group (a significant difference), it would be problematic to express the mean E–C difference in terms of the average within-group *SD* for comparison with the WISC Full Scale IQ difference expressed in the same manner, since the E and C groups do not show significantly different variances on the WISC IQ. The best we can do, therefore, is to compare achievement and IQ differences in terms of the E–C mean difference divided by the *SD* of the E group. When this was done for the test data obtained in grades 1 through 4, the overall average E–C difference in scholastic achievement (i.e., Reading + Math) turned out to be just one-third as large as the overall average E–C difference in IQ.

were 17.75 to 9.39.[2] Although the Math percentiles are lower than those for Reading, especially in the E group, the overall E–C difference for Math is statistically significant ($p < .05$). Without any information on the *SD*s in the C group, it is impossible to tell to what extent this statistical outcome (i.e., $p < .05$) resulted from an artifactual reduction in C group variance, since some of the C subjects are very close to the "floor" on the Math tests or have "bottomed out" beyond the first grade, as Garber notes (p.267).[3] But by the same token, many of the E group have also bottomed out in Math by the end of the fourth grade. It is also noteworthy that by the fourth grade the E group is at the 18.28 percentile on the least *g*-loaded Math test (Computation) and at the 9.79 percentile on the most *g*-loaded Math test (Problem Solving). (The 10th percentile on the WISC corresponds to an IQ of 81 in the normative population.)

At least one conclusion seems fairly certain from these achievement results: the so-called "inoculation" theory is not supported, that is, the idea that the effects of early environmental intervention with children at risk for low IQ will prevent a decline in IQ and its scholastic correlates long after the period of special intervention has concluded.

One possible interpretation of these results, probably favored by the investigators, is that the E children's leaving the Stimulation Center and returning to their poor home environments and low IQ mothers left the E children with too little intellectual stimulation and motivational support (beyond what the public schools provided) to maintain the initial advantage of their enhanced capabilities afforded by the intensive intervention program during their first 6 years of life.

Given the present evidence, a more plausible interpretation, it seems to me, is that the intervention did not materially increase the essential *g* factor that IQ is ordinarily supposed to reflect in terms of the rank order of individuals' levels of *g*. The observed E–C increment in mean IQ most likely represents a non-*g* inflation of test scores that has little, if any, generalization to cognitive achievements that are further removed on the transfer gradient from the specific skills trained by the intervention. I find

[2] The percentile figures quoted here were taken from Table 9–8 (p. 269); Dr Garber has since sent me 9 pages of errata, which include 15 numerical changes in the total of 48 means listed in the book's Table 9–8. With reference to the figures quoted above, the 10.63 has been changed in the errata to 11.94 and the 9.39 to 9.63. Those who wish to perform secondary analyses of the statistics presented in this book are urged to obtain the errata from the author or publisher, as they make corrections in what are apparently numerous (but usually small) errors in six of the important tables, in addition to providing *SD*s for some variables for which only the group means were reported in the book.

[3] Dr. Garber has since provided the needed *SD*s, which bear out my surmise. They show that the overall average *SD* of the percentile scores on Math achievement for the C group is 11.83 as compared with 21.44 for the E group, a significant difference ($F(16,17) = 3.28$, $p < .01$).

many details of the evidence in this report that are consistent with this hypothesis and none that could refute it.

Perhaps more disturbing than the achievement test scores per se is the information in the individual "case histories" of each of the E and C children given in the book's penultimate chapter. A great many of the E as well as the C group apparently turned out to be "adjustment problems" in school by the time they reached the fourth grade—not just academically but in various kinds of "problem behavior," often calling for the attention of school psychologists or other special services. Reading the individual case reports in this regard, it is rather difficult to see much overall difference between descriptions of the E and C children.

These descriptions reminded me of the well-known study by Scarr and Weinberg (1976), in which 176 black and interracial (i.e., white mother/black father) infants were adopted into white upper middle class families, with adoptive parents who were mostly college graduates whose IQs averaged about 120 and whose own (i.e., biological) children's IQs average about 117. The adoptees were not selected for being at risk; if anything, the opposite was the case. At an average age of about 7 years, the IQ means for the black and the interracial adoptees were 96.8 and 109, respectively. But then a 10-year follow-up study found that the adoptees' IQs (and scholastic achievement) declined as the children got older, although they were reared the whole time in intellectually superior and educationally supportive adoptive homes (Scarr, Weinberg, & Gargiulo, 1987). The investigators summarize the main findings of the follow-up study: "Preliminary results indicate that there is considerable decline in the IQ and educational achievements of the Black and interracial Black adoptees and more social deviance and psychopathology than has been reported in previous adoptive samples" (p. 42). This report throws considerable doubt on the conjecture that the Milwaukee E group's decline in IQ and scholastic achievements after age 6 might have been prevented if, following the intervention, they had been able to live in intellectually superior and educationally supportive middle class families instead of their intellectually impoverished environments.

Finally, the one conclusion to which the account of the Milwaukee Projects leads inescapably, that seems so obvious as to need no interpretation, and remains uncontradicted by any position one may take with respect to the nature/nurture argument is that, in this society, women with IQs of 75 or below rearing children is hardly less than a personal tragedy for their offspring, who generally reenact the pathetic life history of their parents. And it is hardly less than a social calamity as well, considering that the estimated percentage of women (and men) with IQs of 75 and below constitutes at least 4 to 5% of the white population and at least 20 to 25% of the black population, and considering that women in this low IQ

segment of the female population have a higher rate of childbearing than women in all the rest of the population.

Garber's full-scale report of the Milwaukee Project is an impressive work, and he deserves much credit for seeing it through. It is neither his fault nor anyone else's fault that, unfortunately, the findings offer no assurance that such extraordinarily intensive and extensive environmental intervention comes even near to being either a feasible or an effective solution to the personal and social misfortunes to which the Project was addressed with high hopes in 1966.

REFERENCES

Cohn, S. J., Carlson, J. S., & Jensen, A. R. (1985). Speed of information processing in academically gifted youths. *Personality and Individual Differences*, **6**, 621–629.

Ericsson, K. A. (1988). Analysis of memory performance in terms of memory skills. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 4). Hillsdale, NJ: Erlbaum.

Garber, H., & Heber, R. (1977). The Milwaukee Project: Indications of the effectiveness of early intervention in preventing mental retardation. In P. Mittler (Ed.), *Research to practice in mental retardation: Vol. 1. Care and intervention*. Baltimore, MD: University Park Press.

Garber, H., & Heber, R. (1981). The efficacy of early intervention with family rehabilitation. In M. J. Begab, H. C. Haywood, & H. L. Garber (Eds.), *Psychosocial influences in retarded performance: Vol. II. Strategies for improving competence*. Baltimore, MD: University Park Press.

Garber, H., & Heber, R. (1982). Modification of predicted cognitive development in high-risk children through early intervention. In D. K. Detterman & R. J. Sternberg (Eds.), *How and how much can intelligence be increased?* Norwood, NJ: Ablex.

Heber, R., & Garber, H. (1975). The Milwaukee Project: A study of the use of family intervention to prevent cultural-familial mental retardation. In B. Friedlander *et al.* (Eds), *Exceptional infant: Vol. III. Assessment and intervention*. New York: Brunner/Mazel.

Heber, R., & Garber, H. (1980). Prevention of cultural-familial mental retardation. In A. Jaeger & R. Slotnick (Eds.), *Community mental health: A behavioral–ecological perspective*. New York: Plenum.

Heber, R., Garber, H., Harrington, S., Hoffman, C., & Falender, C. (1972). *Rehabilitation of families at risk for mental retardation: A progress report* (for the Social and Rehabilitation Service, Department of Health, Education and Welfare, Washington, DC). Madison, WI: Univ. of Wisconsin.

Jensen, A. R. (1982). Reaction time and psychometric *g*. In H. J. Eysenck (Ed.). *A model for intelligence*. Heidelberg: Springer-Verlag.

Jensen, A. R. (1984). Test validity: *g* versus the specificity doctrine. *Journal of Social and Biological Structures*, **7**, 93–118.

Jensen, A. R. (1985a). The nature of the black–white difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences*, **8**, 193–219.

Jensen, A. R. (1985b). Methodological and statistical techniques for the chronometric study of mental abilities. In C. R. Reynolds & V. L. Willson (Eds.), *Methodological and statistical advances in the study of individual differences*. New York: Plenum.

Jensen, A. R. (1986). *g:* Artifact or reality? *Journl of Vocational Behavior*, **29**, 301–331.

Jensen, A. R. (1987a). The *g* beyond factor analysis. In J. C. Conoley, J. A. Glover, &

R. R. Ronning (Eds.), *The influence of cognitive psychology on testing and measurement*. Hillsdale, NJ: Erlbaum.

Jensen, A. R. (1987b). Further evidence for Spearman's hypothesis concerning black–white differences on psychometric tests. *Behavioral and Brain Sciences*, **10**, 512–519.

Jensen, A. R., Cohn, S. J., & Cohn, C. M. G. (1989). Speed of information processing in academically gifted youths and their siblings. *Personality and Individual Differences*, **10**, 29–34.

Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences*, **3**, 423–438.

Page, E. B. (1972). Miracle in Milwaukee: Raising the IQ. *Educational Researcher*, **1**, 8–15.

Page, E. B. (1973). Physical miracle in Milwaukee? *Educational Researcher*, **2**, 2–4.

Page, E. B. (1986). The disturbing case of the Milwaukee Project. In H. H. Spitz (Ed.), *The raising of intelligence A selected history of attempts to raise retarded intelligence*. Hillsdale, NJ: Erlbaum.

Page, E. B., & Grandon, G. M. (1981). Massive intervention and child intelligence: The Milwaukee Project in critical perspective. *Journal of Special Education*, **15**, 239–256.

Plomin, R., & Daniels, D. (1987). Why are children in the same family so different from one another? *Behavioral and Brain Sciences*, **10**, 1–60.

Scarr, S. (1985). Constructing psychology: Making facts and fables for our times. *American Psychologist*, **40**, 499–512.

Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype–environment effects. *Child Development*, **54**, 424–435.

Scarr, S., & Weinberg, R. A. (1976). IQ test performance of black children adopted by white families. *American Psychologist*, **31**, 726–739.

Scarr, S., & Weinberg, R. A. (1978). The influence of "family background" on intellectual attainment. *American Sociological Review*, **43**, 674–692.

Scarr, S., Weinberg, R. A., & Gargiulo, J. (1987). Transracial adoption: A ten-year follow-up [Abstract]. *Program of the 17th Annual Meeting of the Behavior Genetics Association, Minneapolis, Minnesota, June 24–27, 1987.*

Sommer, R., & Sommer, B. A. (1983). Mystery in Milwaukee: Early intervention, IQ, and psychology textbooks. *American Psychologist*, **38**, 982–985.

RECEIVED: September 23, 1988; REVISED: November 1, 1988.