INTELLIGENCE **12**, 315–332 (1988)

Racial Differences in IQ: Item Analysis of the Stanford-Binet at 3 Years

JEANNE E. MONTIE
JOSEPH F. FAGAN, III
Case Western Reserve University

The present study explored the nature of differences in performance on the 3rd revision of the Stanford-Binet for groups of black and white preschoolers matched for parental education in two independent experiments. Large mean differences, favoring the white children, were found in both experiments. In addition, significant race × items interactions at level III, in both experiments, and level III-6, in Experiment 2, indicated that the differences in performance between blacks and whites were much larger on some items relative to others. Results were further examined by contrasting items on which black and white performance was not significantly different with items which showed large significant differences in performance.

The purpose of the present study was to examine the nature of differences in performance of 3-year-old black and white children on the Stanford-Binet Intelligence Scale Form L-M (Terman & Merrill, 1960). Specifically, it was asked whether black children would show a general deficit compared to white children over all items tested or whether racial differences would be disproportionately associated with particular test items.

Racial differences in scores on standard intelligence tests in school age and adult populations have been well documented (see reviews by Shuey, 1966, and Jensen, 1980). In general, when comparing samples of blacks and whites, differences in mean IQ scores of about 15 points, or approximately one standard deviation, are found, consistently favoring the white population (Reynolds & Gutkin, 1981). It is not clear, however, how early in life differences in measures of intelligence between black and white samples can be obtained. Although the issue of whether or not it is possible to measure intelligence in infancy is by no means settled, what evidence exists suggests that groups of black and white infants tested within the first year of life do not differ on measures of mental function. Bayley (1965) found no differences between black and white infants

The research presented in this paper was supported in part by a Mental Retardation Research Training Grant (HD-07176) and by a Major Research Project Grant (HD-11089), both from the National Institute of Child Health and Human Development. The authors wish to thank Dr. Maureen Hack for allowing access to her data on children of very low birth weight.

Correspondence and requests for reprints should be sent to Joseph Fagan, Department of Psychology, Case Western Reserve University, Cleveland, OH 44106.

tested on the Bayley Scale of Mental Development (Bayley, 1969) between 1–15 months. Fagan and Singer (1983) have also noted that black and white infants do not differ significantly on tests of visual recognition memory at 7 months of age, although the same visual recognition tests at 7 months were predictive of measures of intelligence at 3 years within each racial group.

Results of studies carried out with children at the preschool level, with few exceptions, show differences in intelligence test scores between groups of black and white children of about 12–15 points. In her review of studies carried out prior to 1963, Shuey (1966) concluded that the average IQ of the black preschool child is about 12 points below that of the white child the same age. She further noted that the racial differences in IQ scores are decreased when black children are selected for testing on the basis of their living in the same neighborhood, attending the same school, and having fathers in the same occupational groups as their white counterparts.

More recently, Broman, Nichols, and Kennedy (1975) have reported results from the Collaborative Perinatal Project on 12,210 white and 14,550 black 4-year-old children given the short form of the Binet. When socioeconomic status (SES) and sex were controlled, mean IQ differences ranged from about 8 points at the lowest SES category to 13 points at the highest level. In general, when variables such as sex and father's occupation are held constant, significant mean differences in IQ between black and white children are found, favoring the white children. The magnitude of the difference is not consistent, varying with the children's age and the test given. One study (Kaufman, 1973) found the IQ difference to be largest for the youngest children, while others (Arinoldo, 1981; Kaufman & Kaufman, 1973) found IQ differences between blacks and whites to increase with age.

The reasons for the differences between blacks and whites in scores on intelligence tests at later ages have been a subject of much debate in the professional and lay literature. The disparity in black and white IQ scores has been attributed variously to the effects of genetics, socioeconomic status, cultural values, language, and test bias. The present study did not address the nature/nurture issue specifically, but sought to examine the performance of black and white preschool children in an attempt to discover at how early an age differences in performance on a standard intelligence test between black and white children can be documented. A second goal was to note if the pattern of responses to items was similar across races. We asked whether test items would discriminate equally between races or whether blacks would find some items disproportionately more difficult than whites.

Investigations of the issue of internal test bias, that is, the possibility that test items may be disproportionately difficult in different populations, can be carried out a number of different ways. Factor analytic techniques can be employed in order to determine whether or not a test measures the same ability or abilities in each group. A significant group difference in factor loading or factor structure is an indicator of bias (Jensen, 1980). A second technique that can be employed to examine the possibility of bias is that of analysis of variance, with particular attention directed to a group × items interaction. A group × items interaction exists when all the items in the test do not maintain the same relative level of difficulty in each of the groups tested. Jensen (1980) suggested that such an interaction between groups and items can be examined through analysis of variance or through correlation of rank order of item difficulty across groups. He maintained that both methods yield equivalent results when applied to the same set of data.

Miele (1979), for example, looked for possible cultural bias in the WISC (Wechsler, 1949) using analysis of variance to look at the race × items interaction and rank order correlation of item difficulty between races. Two hundred seventy-four children were tested at age 6 years and retested at the end of grades one, three, and five. The sample was balanced for race and sex. A series of race × items analyses of variance revealed significant effects for race and items, and found significant race × items interactions at all age levels. While the significant ($p < .0001$) race × items interactions would indicate that some items were disproportionately difficult in the group tested and therefore be evidence of internal bias, Miele discounted this indication of test bias by pointing out that the amount of variance accounted for by the race × items interaction was never above 5%, while the items effect accounted for over half of the variance at each age level. It is not surprising that the large majority of the variance in performance would be accounted for by the items, given the number of items (161) and range of item difficulty found on the WISC. However, that fact makes it no less significant that the analysis showed differences in the relative difficulty of items between the two groups tested, as indicated by the significant race × items interaction found at all age levels. In a further analysis of internal test bias, Miele rank ordered the 161 items according to percent passing for each group at each age level and rank order correlations were carried out. The mean cross-racial correlations were .96 for males and .95 for females, leading Miele to conclude that the rank order of item difficulty was essentially the same for both groups and, therefore, the WISC does not show evidence of bias for the groups tested.

A similar analysis employing rank order correlation of item difficulty between groups was carried out using the Stanford-Binet. In an investigation of the effects of social class, Goldstein, Meyer, & Egeland (1978) administered the Stanford-Binet to a group of 92 Head Start children, the majority of whom were black, and 92 middle-class white children of the same mental age as the blacks. The 24 items of the Stanford-Binet between years IV and VI were rank ordered according to difficulty within each group. The rank order correlation over item difficulty between groups was .92. Based on their analysis Goldstein et al. concluded that the individual items were of the same relative difficulty in each group. The implication that follows from the correlational analysis is that the test shows no evidence of internal bias for the groups tested.

It would appear from the various correlational analyses performed by Miehle and Goldstein et al. that the high rank order correlations of item difficulty between two groups demonstrate a lack of internal test bias and reflect equivalent patterns of cognitive abilities. It can be demonstrated, however, using a hypothetical example that rank order correlations are spuriously inflated when based on tests in which items vary in difficulty due to age effects. The inflation is a result of failure to separate the influence of group × age effects from age effects. Table 1 presents an example of randomly generated data for two hypothetical groups. In the example, order of "item difficulty" has been randomly assigned over six items within each of four "test levels" for a total of 24 items. Each level represents a more difficult group of items than that found at the previous level. The item difficulty scores for each group are obtained from a random number table with the following restrictions. Scores at level II-6 (i.e., 2 years-6 months) the easiest level, must range from 1–6. At level III, representing slightly more difficult items, values range from 7–12, at level III-6 ranks are 13, 14, 16, 17, 18, and 19, and, at level IV, ranks are 15, 20, 21, 22, 23, and 24. Ranks reflect floor and ceiling effects obtained when test items at year levels II-6, III, III-6, and IV on the Stanford-Binet are rank ordered for difficulty based on McNemar's (1942) analysis of the performance of 3-year-olds over the same items for the 1937 standardization of the Stanford-Binet. Clearly, the easiest items are at the youngest level and the most difficult items at the oldest level measured.

It is instructive to note from the hypothetical data listed in Table 1 that the correlation between items within each difficulty level is never very large, ranging from −.03 to .27 (mean r = .13), which would mean that at any particular level the pattern of item difficulty is not the same for the two groups, thus providing evidence of internal test bias. However, when all 24 items are rank ordered and correlated between groups, the resultant correlation is .92. The value of .92 obtained from our hypothetical example is similar in magnitude to the values of

TABLE 1
Hypothetical Data Representing Rank Order of Item Difficulty for Each of Two Groups at Four Test Levels

Levels		II-6		III		III-6		IV	
Groups		A	B	A	B	A	B	A	B
Items	1.	5	3	7	12	19	16	20	15
	2.	2	2	8	8	17	14	21	23
	3.	1	6	10	9	18	19	15	21
	4.	3	1	12	11	14	13	22	22
	5.	6	5	9	7	16	17	23	20
	6.	4	4	11	10	13	18	24	24
		r = −.03		r = .09		r = .14		r = .27	

.96 and .95 obtained by Miele and equal to the value of .92 reported by Goldstein et al. Perhaps the high correlations obtained by the authors in the previously mentioned studies are the result of a statistical artifact and not indicative of the fact that the items have the same relative difficulty in each group. In any case, it is clear from Table 1 that high rank order correlations of item difficulty across groups can be obtained when the items employed range from very easy to very difficult within each group due, simply, to variations in difficulty with age. In order to avoid the artifactual confounding of age effects with group × age effects, analysis of variance can be carried out over the items within each age level for the groups tested. A significant group × items interaction at a particular age would show that the items were not of the same relative difficulty in each of the groups tested.

In summary, differences in IQ scores between groups of blacks and whites have been well documented in preschool, school age, and adult populations. These differences range from approximately 15 points or one standard deviation in randomly selected samples to much smaller values when samples are matched for sex, geographic location, and socioeconomic variables. It is not clear how early in life these differences in IQ scores between black and white groups can be demonstrated. The little evidence that exists, however, shows no differences between black and white infants tested with the Bayley Scales of Mental Development within the first 15 months of life and no differences in scores on tests of visual recognition memory within the first 7 months of life.

One goal of the present study was to discover how early in life differences in performance on a standard intelligence test could be demonstrated between black and white children matched for age, sex, and parent's education. The Stanford-Binet Intelligence Scale was selected as a measure of intelligence because of its long history of use with very young children. If differences in mean IQ scores between the two groups, favoring the white sample, could be demonstrated, a second goal was to determine whether such differences reflected a general deficit in the performance of the black children across all items or whether some items were disproportionately more difficult for the children in the black sample. Internal test bias due to disproportionate item difficulty levels has traditionally been examined using rank order correlations of item difficulty between two groups or looking for a group × items interaction in an analysis of variance. Since high rank order correlations of item difficulty between two groups may be inflated due to inherent floor and ceiling effects, it was decided to examine the nature of the differences in item difficulty between the black and white samples in the present study using a two-way (race × items) analysis of variance at levels II-6, III, and III-6 of the Stanford-Binet. It was assumed that a significant race × items effect, if obtained, would be evidence that the items were not of the same relative level of difficulty in each group. The two goals of the present study were approached in each of two experiments.

EXPERIMENT 1

Method

Subjects. The subjects included 92 children who were participants in a longitudinal research project studying very low-birthweight infants. The subjects were born between 1975 and 1979 and were selected from a much larger sample on the following basis: birthweight > 1000 g, mothers > 18 years of age, married parents, and no known chronic illness or obvious neurological abnormalities. Forty-six white children and 46 black children were matched for sex and maternal education. The groups did not differ significantly in birthweight, gestational age, or mean age at which the test was given (see Table 2). There were 20 females and 26 males in each group; eight children in each group were outcomes of multiple births.

Birth order did differ between the black and white groups. Excluding children from multiple births, there were 16 first-born and 22 later-born subjects in the black group. In contrast, the white subjects included 27 firstborn and 11 later born. *T*-tests revealed no significant differences in IQ scores between first-and later-born subjects within each racial group. Mean IQ of the first-born blacks = 88.94 (13.14), mean IQ of the later-born blacks = 91.77 (13.14), $t(36) = .68$. Mean IQ of the first-born white = 107.59 (16.86), mean IQ of the later-born whites = 99.91 (11.47), $t(36) = 1.62$.

Procedure. The Stanford-Binet Intelligence Scale was administered to the children as a routine part of the follow-up testing for the research project in which the children were involved. The testing was done by trained examiners, all of whom were white females. The children were tested in their homes or in the pediatric out-patient clinic of a local hospital. The majority of the children were tested at the end of the third year of life (range 28–41 months). Prior to scoring the tests, chronological ages were corrected for prematurity.

Results

The primary aim of the analysis was to determine whether the groups of 3-year-old black and white children differed in performance on the Stanford-Binet and, if so, whether this difference would be in the form of a general deficit of one racial group over all the items, or if group differences would be disproportionately associated with particular test items. Table 2 presents mean IQ scores for the black and white children. A significant difference was found favoring the white children, $t(90) = -4.77$, $p < .001$. The magnitude of the difference is approximately equal to the standard deviation of the test, a value which approximates those found in previous studies comparing unmatched samples of blacks and whites.

Having established that the mean IQ scores of the black and white children differed significantly, an item analysis was carried out in order to determine if

TABLE 2
Means, Standard Deviations, and *t* Values for 46 Black and 46 White Children in Experiment 1

		Black	White	*t*
Maternal education	*M*	4.59	4.80	−.99
	SD	.93	1.09	
Birthweight (in grams)				
	M	1275.75	1262.72	.427
	SD	138.06	152.55	
Gestational Age (in months)				
	M	30.15	29.96	.447
	SD	2.27	1.78	
Chronological Age at Test (in months)				
	M	34.02	33.33	1.34
	SD	2.48	2.47	
Birth Order				
	M	1.79	1.42	2.11*
	SD	.81	.72	
IQ				
	M	90.11	104.09	−4.77**
	SD	12.83	15.21	
	range	68–114	87–147	

*$p < .05$, **$p < .001$

2 = graduate school, 3 = college graduate, 4 = some college, 5 = high school graduate, 6 = some high school, 7 = jr. high school, 8 = < 7 years

the differences in performance between the black and white children tested were of the same magnitude across all items or if the differences in performance varied from item to item. Each of the 24 items on the Stanford-Binet between levels II-6 and IV was scored in terms of percent passing for each racial group. This procedure revealed ceiling effects at level II-6. Blacks and whites did not differ in performance at this level in either sample; between 75% and 100% passed each item. At level IV, floor effects were apparent for both racial groups. It was decided, therefore, to concentrate further analyses on levels III and III-6 where items of intermediate difficulty could be found.

Figure 1 illustrates the percentage passing items 1–6 on the Stanford-Binet at levels III and III-6. It is readily apparent that the white children performed better than the blacks over most items, in addition, it is apparent that the magnitude of the difference in performance varies across items. A two-way (race × items) ANOVA with repeated measures on items was carried out separately at levels III

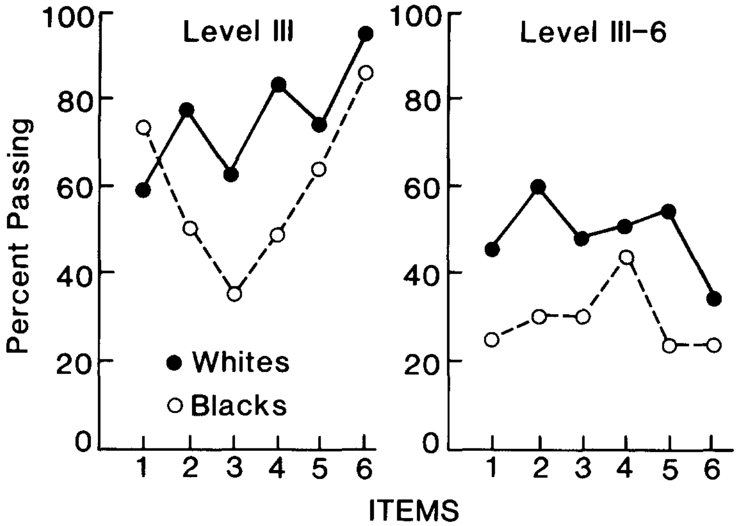

FIG. 1. Percentage of black and white children in Experiment 1 passing items 1–6 on the Stanford-Binet at levels III and III-6.

and III-6. At level III significant main effects were found for race, $F(1,90) = 8.50$, $p < .001$ and items, $F(5,450) = 10.97$, $p < .001$. A race × items interaction also reached significance, $F(5,450) = 4.87$, $p < .001$. Inspection of the graph for level III reveals that, in general, blacks and whites tend to perform in a similar manner on items 1, 5, and 6; while, on items 2, 3, and 4, black performance is far below that of the white group. In other words, the differences in performance are disproportionately associated with particular test items.

A similar analysis was carried out for items at level III-6. Analysis of variance again revealed a significant main effect for race, $F(1,90) = 8.81$, $p < .001$. The main effect for items did not reach significance, $F(5,450) = 2.17$, nor did the race × items interaction, $F(5,450) = 1.24$. It was felt the lack of a significant effect for items and the lack of a significant race × items interaction at level III-6 was due to the fact that floor effects began to be apparent.

As a more stringent test of the possible source of the differences between the two groups, analyses of covariance were performed at each level. Covariates were maternal education, gestational age, birth order, and birthweight; none were statistically significant at level III or at level III-6.

EXPERIMENT 2

Experiment 1 provided evidence that differences in performance between blacks and whites were disproportionately associated with particular test items at level III. However, as performance on any individual test item is not particularly reliable, we felt it was necessary to attempt to replicate the results of Experiment 1 using a new sample. The purpose of Experiment 2 was to attempt to replicate the results of Experiment 1 at level III and obtain more meaningful results at level III-6 by repeating the experiment with a slightly older sample.

Method

Subjects. The subjects included 40 black and 40 white children recruited from county birth lists and local preschools and daycare facilities. All of the subjects resided in lower-middle and middle-class suburbs of Cleveland and were within normal limits for birthweight and gestational age. The groups did not differ significantly in chronological age at test, birth order, or level of parental education (see Table 3). The mean chronological age of the sample was 37.3 months. Each racial group included 21 females and 19 males. There were 12 firstborn and 28 later born in the white sample and 16 firstborn and 24 later born in the black sample.

Procedure. The Stanford-Binet Intelligence Scale was administered to the children either in their homes, preschools, or daycare. The tests were administered by one of three different white female examiners.

Results

The aim of the analysis, as in Experiment 1, was to determine whether the groups of black and white children differed in performance on the Stanford-Binet and, if so, whether this difference would be in the form of a general deficit of one racial group over all the items, or if group differences would be disproportionately associated with particular test items. Mean IQ scores for the children in Experiment 2 are presented in Table 3. As in Experiment 1, a significant difference was found favoring the white children; the magnitude of the difference was slightly less than the standard deviation of the test.

An item analysis was carried out at levels III and III-6 in order to determine if the differences in performance between the black and white children were of the same magnitude across all items or if the differences in performance varied from item to item. It is obvious from Figure 2 that, although the white children performed better than the black children over most items, the magnitude of the difference in performance varied greatly across items. A two-way (race × items) ANOVA with repeated measures on items carried out separately at levels III and III-6 revealed significant effects for race, $F(1,78) = 18.35, p < .001$ at level III, $F(1,78) = 21.52, p < .001$ at level III-6 and items, $F(5,390) = 12.82, p < .001$ at level III, $F(5,390) = 12.42, p < .001$ at level III-6. In addition, at each level a race × items interaction proved significant, $F(5,390) = 9.31, p < .001$ at level III and $F(5,390) = 4.71, p < .001$ at level III-6.

Analyses of covariance were carried out for covariates maternal education and

TABLE 3
Means, Standard Deviations, and *t* Values for 40 Black Children and 40 White Children in Experiment 2

		Black	White	*t*
Maternal education (in years)				
	M	14.23	13.88	.75
	SD	2.12	2.07	
Chronological age at test (in months)				
	M	37.3	37.23	.38
	SD	.85	.92	
Birth order				
	M	2.13	2.35	−.75
	SD	1.25	1.42	
IQ				
	M	97.13	111.20	−5.34**
	SD	13.05	10.36	
	range	69–132	93–139	

**$p < .005$

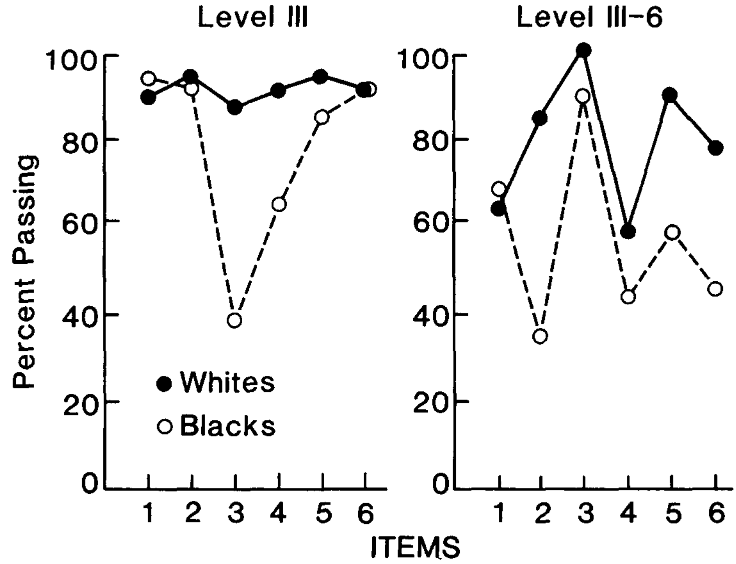

FIG. 2. Percentage of black and white children in Experiment 2 passing items 1–6 on the Stanford-Binet at levels III and III-6.

birth order at levels III and III-6. At level III neither covariate reached statistical significance; at level III-6 birth order was not statistically significant. Maternal education was a marginally significant covariate, $F(1,77) = 3.53$, $p = 0.06$, however, when maternal education was covaried effects for race, items, and race × items remained highly significant ($p < .001$ in each case).

Inspection of the graphs in Figures 1 and 2 reveals that in both samples, at level III, blacks and whites perform in a similar manner on items 1, 5, and 6, while black performance is far below white performance on items 3 and 4. At level III-6 in Experiment 2 blacks and whites perform in a similar manner on items 1, 3 and 4, while black performance falls far below white performance on items 2, 5, and 6. As in Experiment 1, the differences in performance are disproportionately associated with particular test items.

Combined Analyses. $\hat{\Sigma}$-squared values were computed for the analyses of variance at levels III and III-6 for the samples from Experiments 1 and 2 to determine the proportion of the total variance accounted for by each source. Main effects accounted for between 3–14% and 1–10% of the variance for race and items, respectively. The race × items interaction accounted for 3–7% of the variance. Thus, the proportion of the total variance attributed to the two main effects, race and items, and the race × items interaction, are in the same general range.

In an effort to further explore the nature of the differences in performance between the black and white children, a series of chi-square tests were carried out to compare, for each of the twelve items at levels III and III-6, the performance of the black and white groups. Table 4 presents the chi-square values computed for each item using the total combined sample ($n = 172$). Inspection of the table reveals group differences exceeding chance on 6 of the 12 items.

One possible explanation for the race × items interactions obtained at level III in both samples and level III-6 in sample 2 might be that items which are easiest for the white children are those that show no group differences, while large discrepancies in performance occur on items which are the most difficult for the white children. Examination of the data shows that this is not the case. For the items which show no differences between the groups, the percent of white children in both samples passing, averaged over items, = 73%. For those six items that show the largest black–white differences the averaged percent of white children passing is 75%. These numbers indicate that it is not the case that the black children do relatively more poorly on the most difficult items, but perform equally well on those items that are easiest for the white children. In short, one cannot point to the absolute level of difficulty of the items at year III and III-6 to account for the observed race × items interactions.

As an additional analysis, factor analyses were performed to test the hypothesis that differences in the IQ scores of black and white children may be systematically related to the general or g factor in intelligence. In effect, we wanted to

TABLE 4

Percent Passing, Chi-square Values, and *g*-loadings for Each Item at Years III and III-6 on the Stanford-Binet Computed for the Combined Samples in Experiments 1 & 2 ($n = 172$)

		Year III		
Item	black *p*	white *p*	χ^2	*g* loading
1. Stringing beads	84	73	2.79 (ns)	.385
2. Picture vocabulary	70	86	6.62 ($p < .02$)	.604
3. Block buildings: bridge	36	73	24.02 ($p < .001$)	.407
4. Picture memories	55	87	22.11 ($p < .001$)	.488
5. Copy a circle	73	84	2.79 (ns)	.554
6. Draw a vertical line	88	94	1.83 (ns)	.292
		Year III-6		
1. Comparison of balls	45	54	1.14 (ns)	.500
2. Patience: pictures	33	71	25.36 ($p < .001$)	.443
3. Discrimination of animals	58	72	3.69 (ns)	.690
4. Response of pictures: I	44	54	1.89 (ns)	.433
5. Sorting buttons	40	71	17.14 ($p < .001$)	.596
6. Comprehension I	34	55	7.64 ($p < .01$)	.613

know if those items on the Binet showing the highest loadings on *g* would also be the items on which blacks and whites differed most widely in performance. Such an analysis was suggested to us by the studies of Jensen & Reynolds (1982) who found statistically significant correlations ranging from .54 to .75 between black–white subtest differences in the national standardization sample of the WISC-R and a *g* factor derived from the WISC-R battery and Jensen (1985) who found an overall correlation of .59 between *g* loadings and black–white differences for 121 tests in 11 different studies.

In order to increase reliability, the factor analysis was performed on the combined samples 1 and 2, yielding a total sample size of 172. Data at levels III and III-6 in samples 1 and 2 combined were subjected to a principal components analysis. The first unrotated principal component, which accounted for 32% of the variance, was interpreted as a *g* factor. Principal components analysis was also performed on the combined black samples and the combined white samples separately ($n = 86$ in each group). Such an analysis made it possible to compute the degree of similarity between the general factors derived from the white sample, the black sample, and the total sample. The Burt-Tucker coefficient of congruence (Cattell, 1978), r_c was computed yielding a value of .96 for the black *g* factor compared with the white *g* factor, .98 for the black *g* factor compared with the combined *g* factor, and .99 for the white *g* factor compared with the combined *g* factor. Thus, the first unrotated principal component was highly comparable for the total white sample, the total black sample, and the total black and white samples combined.

Black–white difference scores were computed in two ways: as the percentage of black children passing each of the 12 items subtracted from the percent of white children passing each item and as difference scores based on *z* units computed according to the method described by Jensen, 1980[1]. Table 4 presents *g* loadings and percentage passing (*p*) values for the black and the white children in Experiments 1 and 2 over the twelve items. Black–white difference scores, both in terms of percent passing and in terms of *z* units, were then correlated over all 12 items at levels III and III-6 with estimated *g* loadings for each of the 12 items. The resulting Pearson's product–moment correlations were .25 for the percent passing difference and .30 for the *z* difference (rank–order correlations were .30 for percent passing difference and .22 for *z* difference). It is important to note that these values must be interpreted with caution due to the relatively small sample employed and the inherent unreliability of the individual test items.

GENERAL DISCUSSION

In the present study, two independent samples of 3-year-old black and white children matched for sex and parental education within each sample differed significantly in mean IQ scores by approximately one standard deviation. The differences were not simply due to a general deficit in performance by the black children over all the items, but differences in performance between blacks and whites were much larger on some items relative to others, as indicated by the significant race × items interactions in both Experiments 1 and 2 and level III and in Experiment 2 at level III-6. Moreover, items showing large differences between the two racial groups and those items on which the groups did not differ

[1] Z values are used to transform percent passing (*p* values) to an interval scale of difficulty. The difficulty of any given item on the *z* scale is simply the value of *z* on the baseline of the normal curve that is intersected by the vertical line that divides the curve into *p*, the proportion of the total area under the normal curve that passed the item, and $q = 1 - p$, the proportion that failed the item. An item of average difficulty would ($p = .50$) would correspond to a *z* value of 0. Easier items would have *z* values < 0, more difficult items would have *z* values > 0 (Jensen, 1980).

were of the same average difficulty level for the white children, that is, the race × items interactions were not an effect due to black and white children performing equally well on ''easy'' items, but the black children doing relatively more poorly on ''difficult'' items.

The results of the present study are consistent with those reported by Shuey (1966), Broman et al. (1975), Kaufman (1973), and others showing differences in mean IQ scores of approximately one standard deviation or more between samples of black and white preschoolers. In addition, the significant race × items interactions obtained in the present study are consistent with results reported by Miele (1979) in his analysis of performance of black and white grade-school children on the WISC. In the present study, black–white difference scores showed low to moderate correlations with a *g* factor derived from the present samples. In older samples, abundant evidence exists which suggests that black–white differences on a variety of intelligence tests are more strongly related to *g* (Jensen and Reynolds, 1982; Jensen, 1985; Naglieri & Jensen, 1987). It is possible, given the small sample employed, that the *g* factor obtained in the present study is not an accurate representation of *g* or that the relationship of black–white differences in performance to the *g* factor increases with age.

Few studies are available which have examined black–white differences on individual items on the Stanford-Binet at the preschool level. Kamii and Radin (1969) compared the performance of disadvantaged black preschool children with white children from the normative sample of the 1960 revision of the Stanford-Binet matched for mental age. Nichols (1972) provided data on very large samples of black and white children tested with the Stanford-Binet. Nichols found a Stanford-Binet IQ difference of 15 points in favor of the white children, however, because the children were tested at 4 years of age on the short form of the Binet, a meaningful comparison of item analyses cannot be made with the present study. In general, we found the most straightforward approach to interpretation of the race × items interactions in the present study was to contrast items on which black and white performance was not significantly different with items which showed the large and significant differences in performance. Although it is dangerous to make inferences on the basis of performance on individual test items because the reliability of the performance on a particular item is not high, in the present study two independent samples showed similar patterns of performance over items at level III. In addition, converging evidence is provided by the performance of subjects in the Kamii and Radin study. The results reported by Kamii and Radin, though not identical to the pattern of differences in the present study, are quite similar because in both studies the performance of the black children was better than or not significantly different from that of the whites on stringing beads, drawing a line, copying a circle, and discriminating among animals. Furthermore, in both studies performance of the black children was significantly worse than that of the white children on block building: bridge, patience pictures, picture memories, and comprehension.

Bead stringing, copying a circle, and drawing a vertical line, items on which performance by black and white children was not significantly different both in the present study and that of Kamii and Radin, are quite obviously visual-motor in nature. Two of the items showing the largest differences in performance in the present study, patience pictures and block building: bridge, are also visual-motor in nature but involve an added component of mental manipulation of spatial relationships. The source of difficulty for the black children on these test items did not appear to be imitation, they had no trouble imitating bead stringing or drawing a line, nor did it appear to involve manual dexterity. The greatest difficulty seemed to lie in the area of mental manipulation or visualization of spatial relationships.

Two of the items, picture vocabulary and comprehension I, each requiring a verbal response, showed intermediate differences in performance between the two groups. It is possible that differences in performance between black and white children on items requiring verbal responses reflected a lack of knowledge of the concepts required to respond correctly, or simply, a lack of familiarity with the manner in which the questions were phrased and/or a hesitancy to answer due to the strange testing situation. It is possible that communication between the white examiners and the black children was not as clear as that between the examiners and the white children. However, empirical studies looking at racial examiner effects do not support this hypothesis (Graziano, Varca, & Levy, 1982; Sattler, 1982). (It should be noted that few of the studies reviewed have investigated effects at the preschool level.) It is more difficult to make inferences regarding the remaining five items. Two, discrimination of animal pictures and response to pictures, did not show significant differences in performance between black and white children in either the present study or the Kamii & Radin study. Performance on comparison of balls was equal in the present study, but not in the 1969 Kamii & Radin study, a fact which may be due to the intervening influence of television programs, such as Sesame Street, with their emphasis on the concepts big and little. Large discrepancies in performance were also found in the present study on picture memories, an item tapping attention and recognition memory, and button sorting, an item requiring categorization.

It seems clear that the differences in performance are not due either to inability to imitate or to poor fine motor dexterity on the part of the black children. The black children had the greatest difficulty when given problems which required mental manipulation of spatial relationships. Performance of the black children on items requiring verbal responses also fell significantly below that of their white counterparts. Although in the present study the black and white samples have been matched on parental education, we recognize that there are still likely to be cultural differences between the two racial groups. It is possible, for instance, that cultural differences concerning the value placed on different types of early learning experiences or cognitive styles are responsible for some of the observed differences in the two samples.

Three additional sources provide evidence of black–white differences on measures of intellectual function similar to the differences found in the present study. Analyses of data from the standardization sample of the 4th edition of the Stanford-Binet Intelligence Scale (Thorndike, Hagen, & Sattler, 1986) show the greatest black–white differences to be in the areas of verbal reasoning and abstract/visual reasoning with smaller black–white performance discrepancies in the areas of quantitative reasoning and short-term memory. A second source of support comes from a recent study by Naglieri and Jensen (1987) which compares black–white differences on the WISC-R and the Kaufman Assessment Battery for Children (K-ABC) in school-age populations. Black–white differences larger than would be predicted from the subtest's *g* loadings were found on two K-ABC subtests, Spatial Memory and Triangles, and two WISC-R subtests, Object Assembly and Block Design. Each of these subtests involve some spatial ability factor in addition to *g*, leading the authors to conclude that a spatial ability factor increases black–white differences on these subtests. A third source of support is found in a review of studies reporting a "perceptual defect" in black populations by Mandler and Stein (1977). Although the authors conclude there is insufficient evidence to support the existence of a perceptual defect in the black population, they note the consistent finding over a large number of studies of lower performance by black children on the block design test. The authors conclude that there may be a visualization factor involving rotations of shapes that differs among ethnic groups, but emphasize that well-controlled sophisticated studies have not yet been carried out to more carefully examine this hypothesis.

It is important to note that the design of the present study does not allow us to attribute the black–white differences to a particular source, but it does provide evidence of how early in life black–white differences on standard measures of intelligence occur. In addition, by looking at the converging evidence of similar black–white differences it is possible to zero in on areas for future study. The fact that the greatest black–white differences were found on items requiring mental manipulation, no significant differences were found on visual-motor items and intermediate differences were found on verbal items world seem to be a fruitful area for future investigation. It would be particularly interesting to use the 4th edition of the Stanford-Binet to test samples of 3-year-old black and white children matched for parental education. Using the newest edition of the Binet it would be possible to compare the performance of groups of black and white children matched for parental education on individual items as well as in the composite areas of verbal reasoning, abstract/visual reasoning, quantitative reasoning, and short-term memory. Results of the present study would predict that the greatest black–white discrepancy would be in the area of abstract/visual reasoning, moderate differences would be found in the verbal area, and little or no difference would be apparent in the quantitative and short-term memory areas.

Finally, given that there now exist infant intelligence tests which are predictive of intelligence later in life (Fagan, Singer, Montie, & Shepherd, 1986) and that evidence to date shows no difference in performance between blacks and whites on infant intelligence tests, it is important to trace back when racial differences on intelligence tests first appear. It may be that the infant intelligence tests are tapping an aspect of intelligence (e.g., fluid intelligence) that remains constant throughout life in both races, and that later race–IQ discrepancies are due to differences in crystallized intelligence which can be altered by intervention. If a careful analysis is done to document at what age IQ differences between black and white children first appear and what areas of reasoning are affected, it may be possible to concentrate intervention efforts on specific areas in the early years of life and reduce later black–white IQ discrepancies.

REFERENCES

Arinoldo, C.G. (1981). Black/white differences in the general cognitive index of the McCarthy Scales and in the full scale IQs of Wechsler's Scales. *Journal of Clinical Psychology, 37,* 630–638.

Bayley, N. (1965). Comparisons of mental and motor test scores for ages 1–15 months by sex, birth order, race, geographic location, and education of parents. *Child Development, 36,* 379–411.

Bayley, N. (1969). *The Bayley scales of infant development.* New York: Psychological Corporation.

Broman, S.H., Nichols, P.L., & Kennedy, W.A. (1975). *Preschool IQ: Prenatal and early developmental correlates.* Hillsdale, NJ: Erlbaum.

Cattell, R.B. (1978). *The Scientific Use of Factor Analysis in the Behavioral and Life Sciences.* New York: Plenum.

Fagan, J.F., & Singer, L.T. (1983). Infant recognition memory as a measure of intelligence. In L.P. Lipsitt (Ed.), *Advances in Infancy Research,* Vol. 2. Norwood, NJ: Ablex.

Fagan, J.F., Singer, L.T., Montie, J.E., & Shepherd, P.A. (1986). Selective screening device for the early detection of normal or delayed cognitive development in infants at risk for later mental retardation. *Pediatrics, 78,* 1021–1026.

Goldstein, D., Meyer, W., & Egeland, B. (1978). Cognitive performance and competence characteristics of lower and middle-class preschool children. *Journal of Genetic Psychology, 132,* 177–183.

Graziano, W.G., Varca, P.E., & Levy, J.C. (1982). Race of examiner effects and the validity of intelligence tests. *Review of Educational Research, 52,* 469–497.

Jensen, A.R. (1980). *Bias in Mental Testing.* New York: Free Press.

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

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

Kamii, C.K., & Radin, N.G. (1969). The retardation of disadvantaged Negro preschoolers: Some characteristics found from an item analysis of the Stanford-Binet test. *Psychology in the Schools, 6,* 283–299.

Kaufman, A.S. (1973). Comparison of the performance of matched groups of black children and white children on the Wechsler Preschool and Primary Scale of Intelligence. *Journal of Consulting and Clinical Psychology, 41,* 186–191.

Kaufman, A.S., & Kaufman, N.L. (1973). Black–white differences on the McCarthy Scales of Children's Abilities. *Journal of School Psychology, 11,* 196–206.

Mandler, J.M., & Stein, N.L. (1977). The myth of perceptual defect: Sources and evidence. *Psychological Bulletin, (84),* 173–192.

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

Miele, F. (1979). Cultural bias in the WISC. *Intelligence, 3,* 149–164.

Naglieri, J.A., & Jensen, A.R. (1987). Comparison of black–white differences on the WISC-R and the K-ABC: Spearman's Hypothesis. *Intelligence, (11),* 21–43.

Nichols, P.L. (1972). The effects of heredity and environment on intelligence test performance in 4- and 7-year-old white and Negro sibling pairs. Doctoral dissertation, University of Minnesota.

Reynolds, C.R., & Gutkin, T.B. (1981). A multivariate comparison of the intellectual performance of blacks and whites matched on four demographic variables. *Personality and Individual Differences, 2,* 175–180.

Sattler, J.M. (1982). *Assessment of children's intelligence and special abilities* (2nd ed.). Boston: Allyn & Bacon.

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

Terman, L.M., & Merrill, M.A. (1960). *Stanford-Binet intelligence scale.* Boston: Houghton Mifflin.

Thorndike, R.L., Hagen, E.P., & Sattler, J.M. (1986). *Stanford-Binet Intelligence Scale: Fourth Edition, Technical Manual.* Chicago: Riverside Publishing.

Wechsler, D. (1949). *Wechsler intelligence scale for children.* New York: Psychological Corporation.