PERGAMON

Personality and Individual Differences 26 (1999) 373–379

PERSONALITY AND INDIVIDUAL DIFFERENCES

Evidence against Rushton: The genetic loading of WISC-R subtests and the causes of between-group IQ differences

James R. Flynn*

Department of Political Studies, University of Otago, P.O. Box 56, Dunedin, New Zealand

Abstract

Rushton ranks WISC-R subtests both for genetic loading and the magnitude of the black–white score gap. He finds a positive correlation and therefore, infers that the genetic contribution to the black–white IQ gap is robust. Rushton's method was applied to five independent data sets showing IQ gains over time. Because these gains are known to be environmental in origin, it should have given robust negative correlations. In fact: for the totality of Wechsler subtests, it implies a nil correlation; for verbal subtests, it gives positive correlations ranging from 0.300 to 0.900; for performance subtests, it gives a mix of positive and negative.

1. Introduction

The review in this journal of Gould, *The Mismeasure of Man*, by Rushton (1997) makes many valid critical points. I would add that Gould's book evades all of Jensen's best arguments, for a genetic component in the black–white IQ gap, by positing that they are dependent on the concept of *g* as a general intelligence factor. Therefore, Gould believes that if he can discredit *g*, no more need be said. This is manifestly false, Jensen's arguments would bite no matter whether blacks suffered from a score deficit on one or 10 or 100 factors. I attribute no intent or motive to Gould, it is just that you cannot rebut arguments if you do not acknowledge and address them.

However, I will present evidence against one of Rushton's points. He says that Gould should be impressed by "the critically important finding that genetic weights on IQ subtests predict racial differences" (Rushton, 1997, p. 176). In his recent book, Rushton (1995, pp. 185–187) places considerable emphasis on this point. He ranks WISC-R subtests in ascending order of black–white

* Tel.: +64-3-479-8668; Fax: +64- 3-479-7174; E-mail: polstuds@gandalf.otago.ac.nz

S0191-8869/98/$—see front matter
PII: S0191-8869(98)00149-4

differences and shows that there is a positive correlation with the genetic loading of the subtests as measured by inbreeding depression. As he says, as the genetic loading increases, so do the black–white differences. He does not provide a value, but the Spearman rank-order correlation comes to +0.531. This has a probability of arising by chance of 0.049, so it has a reasonable level of significance. Rushton (1997, p. 176) believes this method is an important test of differential predictions, a positive correlation verifying a strong genetic component in the racial IQ gap, the failure to find a negative correlation discrediting an environmental explanation. His book (Rushton, 1995, p. 187) claims we must conclude that "the genetic contribution to racial differences in mental ability is robust".

I wish to test Rushton's method on a wider array of data. The ideal evidence is the IQ gap between generations, the result of massive IQ gains over time, because that gap is known to be caused by environmental factors. Rushton and I and others agree that reproductive trends have been dysgenic. So environmental factors account for more than 100% of between-generation IQ differences, having had to overcome a genic negative trend. The only genetic factor seriously proposed as a cause of IQ gains is hybrid vigour, engendered by increased outbreeding. Whatever relevance this may have had because of the rural to urban migration in the Western world beginning about 1800, few would argue that it has significant impact on the data I will present, data from the United States, West Germany, Austria, and Scotland, covering the period from 1947 to 1989. Rushton's method clearly rests on this assumption: whenever between-group IQ differences are primarily genetic, there will be a positive correlation between subtest genetic loadings and the magnitude of between-group score differences on those subtests; whenever the IQ differences are primarily environmental, there will be a negative correlation.

2. Rushton and IQ gains

I have located five data sets in which IQ gains over time have been broken down by WISC-R subtests; others are invited to contribute. United States data cover first, gains from the WISC (normed 1947–1948) to the WISC-R (normed 1972). There are only three studies in which both tests were administered counterbalanced to normal subjects and results for all subtests published (Rowe, 1976; Schwarting, 1976; Stokes et al., 1978). The results for the total of 245 subjects were merged using weighted averages; the WISC-R scores were translated into scores normed on the white members of the WISC-R standardization sample using Jensen and Reynolds (1982, Table 1). The translation is necessitated by the fact that the WISC was normed on whites only. United States data also cover gains from the WISC-R (normed 1972) to the WISC-III (normed 1989) and come from 206 subjects who took both tests counterbalanced, as presented in the WISC-III manual (Wechsler, 1992, p. 198).

There are data from three other nations. Schallberger (1987, p. 9) gave both the West German WISC (normed 1954) and WISC-R (normed 1981) to 124 children counterbalanced. Schubert and Berlach (1982, p. 262) recorded the scores of 2,318 subjects who took the West German WISC between 1962 and 1979, children referred to an Austrian clinic because of scholastic or behavioural problems. They claim their subjects should not have been atypical because such problems were not confined to children with either superior or inferior intelligence. Flynn scored the 10-year olds ($N=155$) and 13-year olds ($N=142$) from the Scottish WISC-R standardisation sample of 1983–

Table 1
Spearman rank-order correlations between genetic loading and score gains over time; WISC-R subtests arranged or numbered in ascending order (1 = lowest) for both variables

WISC-R[a]		Inbreeding depression score	US: WISC to WISC-R		US: WISC-R to WISC-III		West Germany		Austria		Scotland	
subtest	(V/P)		gain	(rank)	gain	(rank)	gain	(rank)	gain	(rank)	gain	(rank)
Cod	P	4.45	2.20	(9)	0.7	(5)	2.8	(6)	1.1	(8)		
A	V	5.05	0.36	(1)	0.3	(2)	−0.5	(1)	−0.4	(1)	−0.9	(1)
M	P	5.35			1.2	(8/9)						
BD	P	5.35	1.28	(7)	0.9	(6/7)	4.8	(9)	1.8	(9)		
PC	P	5.90	0.74	(4)	0.9	(6/7)	4.8	(9)	0.6	(4)		
Com	V	6.05	1.20	(6)	0.6	(4)	2.4	(3/4)	0.1	(2)	3.2	(4)
OA	P	6.05	1.34	(8)	1.2	(8/9)	2.4	(3/4)	1.0	(6)		
I	V	8.30	0.43	(3)	−0.3	(1)	1.0	(2)	0.2	(3)	0.7	(3)
PA	P	9.40	0.93	(5)	1.9	(11)	4.8	(9)	0.9	(5)		
S	V	9.95	2.77	(10)	1.3	(10)	4.7	(7)	2.1	(10)	3.7	(5)
V	V	11.45	0.38	(2)	0.4	(3)	2.6	(5)	1.0	(7)	0.0	(2)
r_s All subtests (Sig):			−0.079	(0.42)	+0.177	(0.30)	+0.074	(0.42)	+0.162	(0.33)		
r_s Verbal subtests (Sig):			+0.300	(0.34)	+0.300	(0.34)	+0.800	(0.07)	+0.900	(0.04)	+0.300	(0.34)
r_s Performance subtests (Sig):			−0.500	(0.23)	+0.806	(0.04)	+0.112	(0.47)	−0.600	(0.18)		

Data and sources: Discussed in text.

Note: The gains on WISC-R subtests are scaled scores based on SD = 3 (rather than 15). Therefore, they must be multiplied by 5 to get something like IQ gains over time. For example, U.S. subjects gained 13.85 points (5 × 2.77) on similarities between 1947–1948 and 1972 (WISC to WISC-R). Even then, comparing nations for rate of gain must take into account the differing time spans the above data cover. For example, for the United States, earlier (WISC to WISC-R) and later (WISC-R to WISC-III) gains on vocabulary look about equal. However, the former took place over 24.5 years and the latter over 17 years, so the later gain is really 1.5 times as great. Also see Flynn (1987, pp. 182–183) for comments on the quality of the West German and Austrian data.

[a] Full names of subtests: coding, arithmetic, mazes, block design, picture completion, comprehension, object assembly, information, picture arrangement, similarities, vocabulary. V designates verbal subtests, P designates performance subtests.

1984 against the WISC norms of 1961–1962, using 68 items that were left unaltered between the two tests. The WISC standardisation sample numbered 2103 including 210 10-year olds and 222 13-year olds. Flynn (1990, p. 47) gives results for five verbal sub-tests which contained a reasonable number of unaltered items. For arithmetic, there were only 5 such items out of 9 or 15 (depending on age) which is barely adequate. But for the other four, well over half the items were unaltered, ranging up to 9 out of 10 for comprehension and 21 out of 26 for similarities.

Rushton's ranking of WISC-R subtests for genetic loading was taken as given, and compared to the same subtests ranked for magnitude of IQ gains. Spearman rank-order correlations were computed for all subtests available, first for the totality of subtests ($N=10$ or 11), second for gains on verbal subtests ($N=5$) and third for performance subtests ($N=5$ or 6). Pearson correlations were not used because of the small number of pairs, and because differential IQ gains on subtests do not show a normal distribution. For example, if quality of schooling is not enhanced over time, you may get no or low gains for arithmetic, information and vocabulary. While non-school factors, some have suggested the advent of video games, may cause large gains on coding, mazes and block design. A real-world illustration: over the last 30 years, black children may well have made large skill gains for well-paid sports like basketball, football and baseball; I would wager they have made no gains on archery, showjumping and small-bore rifle shooting.

It will be recalled that Rushton's method predicts robust negative correlations between genetic loading and magnitude of IQ gains. As Table 1 shows, the data both collectively and by kind of test falsify that prediction. Taking all subtests, four data sets, two U.S. plus West Germany and Austria, give correlations ranging from -0.079 to $+0.177$, for an average of $+0.084$. This amounts to a pattern of no correlation at all. Taking performance subtests, we get correlations ranging from -0.600 to $+0.806$, for an average of -0.046, which is to say that Rushton's method gives no consistent result. It implies that performance gains were: heavily genetic in the U.S. before 1972, heavily environmental thereafter; heavily genetic in Austria, but mildly environmental in West Germany. Taking verbal subtests, five data sets, the fifth is the Scottish data, show correlations ranging from $+0.300$ to $+0.900$, for an average of $+0.520$. That Rushton's method consistently classifies verbal gains as genetic is particularly anomalous, to say nothing of incongruities like performance gains in Austria being heavily genetic, while verbal gains were heavily environmental.

3. Concerning probabilities

The significance values in Table 1 show only two results attaining a reasonable level. For Austrian verbal gains and U.S. (WISC-R to WISC-III) performance gains, the probability of getting the positive correlations ($+0.900$ and $+0.806$) by chance is less than 0.05, in both cases. However, rather than taking data sets independently, we can calculate the probability for a group of sets, say the verbal group, by multiplying the independent probability values. The probability of getting the verbal results by chance are 0.34 times 0.34 times 0.07 times 0.04 times 0.34 times, or 0.00011, eleven chances in 100,000. But this is the probability assuming the null hypothesis, that is, that genetic loading of subtests and IQ gains on subtests are uncorrelated. Since the null hypothesis itself counts against Rushton's method, we have an underestimate. What correlation does Rushton need to inspire confidence in his method? Given that the method gives a positive correlation of 0.531 for the racial IQ gap, which no one argues to be entirely genetic (Jensen puts

the environmental contribution at about one-third), it should give a negative correlation of at least −0.500 for the generational IQ gap, which all concede to be entirely environmental or almost. Otherwise, its results are hardly consistent. However, given that no method is perfect, it might be argued that a negative correlation of −0.300 would do. Certainly, anything less than that is explaining so little variance, less than 9%, as to give no clear lead.

It is easy to calculate values using the null hypothesis because it gives one simple scenario in which all possibilities have an equal weighting. Positing something like −0.500 or −0.300 as the true correlation, and then calculating the probability of say +0.900, is complex because the posited true correlation could arise in a multitude of possible worlds. Fortunately, the computer can simulate almost innumerable worlds and by taking the highest probability engendered, we can get a safe conservative estimate. Using five pairs, the number of verbal subtests, the computer simulated 100,000 samples from several distributions, bivariant normal, bivariant uniform, and bivariant exponential, assuming each time a range of negative values as the true correlation. Assuming −0.500, our verbal results for Austria at +0.900 and West Germany at +0.800 have probabilities of less than 0.01 and the other verbal results at +0.300 have approximately 0.03. Assuming −0.300, the +0.900 and +0.800 have probabilities of less than 0.02, the +0.300s approximately 0.05. Even the −0.300 values, when multiplied, give our verbal results a collective probability of 0.00000005, or five chances in 100 million.

However, this does more to show the power of this way of testing Rushton's method *in potentia*, than it does to estimate the strength of the actual data. If we could be certain of our verbal subtest hierarchies, even with only five subtests, Rushton's method would be overwhelmingly discredited. But these hierarchies represent differential gains on subtests as measured by standardisation samples and the message of the samples is read by taking the tests the samples norm and administering them to groups of subjects. The problem is not so much the standardisation samples. Wechsler samples are of good quality and just so long as they are within a few points of the true population mean, that is, cover essentially normal subjects, they should not distort subtest differentials. The problem is that the experience of measuring differential rates of gain, for full scale, verbal and performance IQ, show estimates firming up after we accumulate about a thousand subjects (who took both of the relevant tests). The numbers here range from 124 to 245 for three of our data sets. The 297 Scottish subjects are more reliable, because they are drawn directly from one of the standardisation samples and exhaust its 10 and 13-year olds and the subtest differentials are large, ranging from 0.167 SDs to a full SD. On the other hand, recall that there were only five unaltered items on the Arithmetic subtest. The Austrian subjects at 2,318 are impressive but recall that while covering the normal range of IQs, they did suffer from scholastic or behavioural problems.

I would prefer to assess the present status of the evidence conservatively, by using the bare fundamentals of probability theory. Every one of our five verbal data sets and every one of our four all-subtest data sets, falsifies Rushton's prediction. Let us assume that they are no more reliable than the flip of a coin, giving one chance in two of verifying, one chance in two of falsifying. The collective probability of the verbal group is then two to the fifth, or one in 32, a probability of 0.03. The collective probability of the all-subtests group is two to the fourth, or one in 16, a probability of 0.06.

The performance group shows two data sets out of four for which Rushton's prediction works. I suspect that sooner or later, it will work for a verbal data set. But it is sobering to reflect how

this might come about. Take the recent U.S. data set covering WISC-R to WISC-III, the years 1972 to 1989. What if in 1989, U.S. schools began to teach arithmetic a lot better and the domination of the out-of-school environment of U.S. children by visual stimuli brought verbal similarities gains to a halt. That alone, all other subtest gains being unaltered, would give us WISC-III to WISC-IV data in which a verbal correlation of +0.300 would become −0.700. For all subtests, the correlation of +0.177 would become −0.493. In sum, because of purely environmental changes, data that falsify Rushton's prediction would suddenly offer it robust confirmation. Or to use a previous example, assume that inbreeding depression affects speed and power more than concentration and steadiness of hand. Then basketball, football and baseball would have higher genetic loadings than archery, show jumping and small-bore rifle shooting. When purely environmental factors caused larger performance gains for the former than the latter, we would appear to have, using Rushton's method, persistent confirmation that sport gains over time were genetic. Can this method ever really inspire confidence as a way of determining whether group differences are caused by environmental or genetic factors?

4. The race and IQ debate

A cautious conclusion: Rushton's method at present can play no persuasive role in the race and IQ debate. Unless the tide of evidence reverses dramatically, it classifies IQ gaps between generations as genetic, or as inconclusive, when these gaps are knows to be environmental. It has a conceptual weakness at its core: when we spell out the real-world causes that might appear to confirm Rushton's method, purely environmental trends engender a hierarchy of Wechsler subtest differences that mimick a hierarchy of genetic loadings. Is it not time to call a moratorium on using this kind of methodology, a methodology that goes from within-group heritability data (used to get subtest genetic loadings) to draw conclusions about between-group IQ differences? Jensen once used low E^2 or environmental estimates (from within-races) to infer that the black–white IQ gap (between-races) had a strong genetic component. Flynn (1989) noted that the same low E^2 estimates held within generations, yet the between-generation IQ gap had no genetic component whatsoever. The argument was shown to be invalid for IQ *globally*, it has now been shown to be suspect for IQ *dissected* into various subtests. The exit of such argumentation from the race and IQ debate might serve as a step towards common ground and clarification.

References

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*, 171–191.

Flynn, J. R. (1989). Rushton, evolution and race: An essay on intelligence and virtue. *The Psychologist*, *2*, 363–366.

Flynn, J. R. (1990). Massive IQ gains on the Scottish WISC: Evidence against Brand *et al.*'s hypothesis. *The Irish Journal of Psychology*, *11*, 41–51.

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

Rowe, H. A. H. (1976). *The comparability of WISC and WISC-R*. Hawthorn, Victoria: The Australian Council for Educational Research.

Rushton, J. P. (1995). *Race, evolution and behavior: A life history perspective*. New Brunswick, NJ: Transaction Publishers.

Rushton, J. P. (1997). Race, intelligence and the brain: The errors and omissions of the 'revised' edition of S. J. Gould's *The Mismeasure of Man* (1996). *Personality and Individual Differences*, *23*, 169–180.

Schallberger, U. (1987). HAWIK und HAWIK-R: Ein empirischer Vergleich [HAWIK and HAWIK-R: An empirical comparison]. *Diagnostica*, *33*, 1–13.

Schubert, M. T., & Berlach, G. (1982). Neue Richtlinien zur Interpretation des Hamburg Wechsler-Intelligenztests für Kinder (HAWIK) [New guidelines for the interpretation of the Hamburg Wechsler Intelligence Tests for Children (HAWIK)]. *Zeitschrift für Klinische Psychologie*, *11*, 253–279.

Schwarting, F. G. (1976). A comparison of the WISC and WISC-R. *Psychology in the Schools*, *13*, 139–141.

Stokes, E. H., Brent, D., Huddleston, N. J., Rozier, J. S., & Marrero, B. (1978). A comparison of WISC and WISC-R scores of sixth grade students: implications for validity. *Educational and Psychological Measurement*, *38*, 469–473.

Wechsler, D. (1992). *WISC-III: Wechsler Intelligence Scale for Children* (3rd ed.) (Australian adaption). New York: The Psychological Corporation, Harcourt Brace Janovich.