Constancy of IQ Scores Among Gifted Children*

Sorel Cahan
Alicia Gejman

In light of the outdatedness of empirical research on IQ constancy among gifted children, and with the aim of examining possible cross cultural differences, the present study investigated the issue within the Israeli context. Specifically, we analyzed the constancy of IQ scores on the WISC-R test for 161 kindergarteners through fourth graders identified as gifted by the Jerusalem Psychological Service in 1981/82 - 1983/84. Assessment of IQ constancy was based on a retest administered to subjects 1-4 years after the first test. Results showed that 86% of the children in the sample were defined as gifted also on retest. Mean absolute differences between testings ranged from 1/3 to 1/2 SD (5-8 IQ points) for Verbal, Performance and Full-scale IQ scores, and from 1/2 to 3/4 SD for subtest scores. On the whole, Performance scores remained constant, while Verbal scores tended to decline. There were no consistent differences attributable to age of identification or measurement interval.

Sorel Cahan is a Senior Lecturer at the School of Education, the Hebrew University of Jerusalem. His work focuses on psychological and educational measurement, especially intelligence and intelligence testing. **Alicia Gejman** is an educational psychologist.

The conceptualization of giftedness among children has changed greatly in the past 50 years, and current definitions recognize multiple talents and abilities (e.g., Feldhusen & Hoover, 1986; Sternberg, 1986; Tannenbaum, 1983; Treffinger & Renzulli, 1986). Nonetheless, intelligence testing still plays a major role in the early selection of students for gifted programs (Silverman, 1986). Early identification is based on the assumption that intelligence test scores remain constant over a considerable period. Indeed, the main objective of the definition of the deviation IQ was to arrive at an intelligence measure that is independent of the subject's age (Thurstone, 1926; Wechsler, 1974). As such, deviation IQ scores are expected to remain constant over the individual's life span (Matarazzo, 1972).

The traditional definition of "constancy" — i.e., the correlation between IQ scores on two examinations taken at different points in time and administered to non-selective populations — is invalid in the context of gifted children. As such children are a selective population, these correlations do not exhaust the meaning of "constancy"; they merely measure the stability of each child's relative ranking within the gifted population. Whereas average IQ deviation scores do remain constant from age to age among the general population, the mean scores of gifted children, who lie at one of the extremes of the population distribution, are liable to change (particularly to decline), and both the direction and magnitude of this change are meaningful aspects of the constancy issue. Thus, the difference between mean scores is a much more revealing measure of constancy among gifted children.

Empirical evidence on IQ constancy among gifted children is scarce mainly owing to the logistic difficulties and high cost of the longitudinal studies that are required. Most empirical research on this topic was conducted early in the century among small sample populations in the U.S. (Cattell, 1933; Hildreth, 1943; Hollingworth, 1938; Hollingworth & Kaunitz, 1934; Lincoln, 1935; Terman, 1925, 1930; Thorndike, 1940; Witty, 1940). Most studies pointed to a downward trend in mean IQ scores at the second examination, although the extent of that decline varied. The decline was greater among girls (Terman, 1930; Lincoln, 1935), among children identified as gifted at a relatively early age (Terman, 1930; Hildreth, 1943) and when the time interval between measurements was relatively long (Lincoln, 1935).

The question of IQ constancy among gifted children holds major importance both from a theoretical and a pragmatic point of view. As previous empirical research is both outdated and specific to the American population, additional studies are needed to examine the universality of existing findings. Moreover, single (one-time) administrations of IQ tests for the purpose of early identification of gifted children are practiced widely. Hence, new evidence regarding constancy of the IQ scores could lead to modifications of the method for identifying gifted children.

The present study examines the size and direction of change in mean IQ scores among Israeli children identified as gifted. It focuses on three main questions: the degree of mean score differences between testings, the direction of that change, and the consistency of change over subtests per subject.

METHOD

Research Population and Sample

The research population included all kindergarteners to fourth graders who were identified as gifted by Jerusalem's Psychological Service in the 1981/82-1983/84 school years for the purpose of acceptance to a special extra-curricular enrichment program. The method used to select elementary-school children for participation in this particular program has been uniform for many years — namely, the individually administered Hebrew version of the Wechsler Intelligence Scale for Children - Revised (WISC-R; Lieblich, Ben Shakhar-Segev & Ninio, 1976). The criterion for defining a child as gifted is either a Verbal or a Performance IQ score that is greater than or equal to 130.

Of the 283 children in the research population, 161 (57%) agreed to participate in the study. These were divided into four research groups, by age of identification and the interval between the two examinations. Specifically, they were split into younger (K-Grade 2) and older (Grades 3 and 4) groups, according to their grade level at the time of identification, and further divided into subjects with short (1-24 months) and

*This study was supported by a grant from the Israeli Ministry of Education and Culture. We would like to thank Lavi Artman for his comments on previous drafts and Helene Hogri for her editorial assistance.

Article submitted February, 1992.
Revision accepted August, 1992.

Means and standard deviations (in parentheses) of IQ scores on the first test, sample (S) and research population (P), by age of identification and measurement interval

	YOUNGER				OLDER			
	Short		Long		Short		Long	
	S (N=33)	P (N=61)	S (N=52)	P (N=80)	S (N=43)	P (N=70)	S (N=33)	P (N=72)
Full-scale	138 (6)	139 (8)	138 (7)	138 (8)	139 (7)	137 (7)	139 (7)	137 (7)
Verbal	138 (8)	138 (8)	138 (10)	137 (10)	135 (7)	134 (8)	136 (6)	134 (8)
Performance	130 (10)	132 (10)	130 (10)	132 (11)	135 (10)	133 (10)	134 (11)	132 (10)

Table 1

long (25-48 months) measurement intervals. Table 1, which gives the number of subjects per research group, as well as the mean IQ scores for each group (Verbal, Performance and Full-scale) at the first examination, shows that there are virtually no differences in IQ scores between research groups. For the sake of comparison, the Table also includes parallel data for the research population. These show that the sample in each group is fairly representative of the research population.

Procedure

Assessment of IQ score constancy was based on a retest of all subjects. Tests were administered individually at the Jerusalem Psychological Service by six experienced educational psychologists over a period of six months during the 1985/86 school year. These examiners were not the same ones who administered the first test.

RESULTS

The Size of Change

Table 2 presents mean absolute differences between initial test and retest for the four research groups, twelve subtests and three IQ scores. For the sake of simplicity, differences in subtest scores are presented in IQ units, such that one point on the subtest scale is equal to 5 IQ points. Mean absolute differences for the three IQ scores ranged from 5 to 8 IQ points (.33-.53 SD). Differences were generally larger for the subtests, ranging from 8 to 11 IQ points (.53-.73 SD), with the exception of the Mazes test, for which differences ranged from 15 to 22 IQ points (1-1.47 SD). No consistent differences were found by age of identification or measurement interval.

For the sake of comparison, we also included in Table 2 the standard error of measurement of the WISC-R scores, computed from the constancy coefficients in Table 11 of the WISC-R manual. These coefficients were obtained by retesting 45 subjects (aged 10 1/2) in the standardization sample 15 days after the initial administration of the test. Thus, these data provide an estimate of the mean absolute differences expected between scores on two near simultaneous administrations of the same test. The mean differences obtained in the present study after an average interval of 2.5 years are larger than the simultaneous estimates by about 50%. The difference is particularly large for the Full-scale IQ. It should be stressed, however, that the estimates of the constancy coefficients in the WISC-R manual are based on a very small sample at a specific age, a fact which considerably affects their generalizability.

The Direction of Change

We examined the direction of change by performing two separate analyses: a comparison between the proportion of negative and positive differences (Table 3) and a calculation of mean non-absolute differences (Table 4). Both tables indicate a general decline in scores for the Verbal subtests (with the exception of "Digit Span") and the Verbal IQ: there were a majority of negative differences between testings and corresponding negative mean differences. The decline was more consistent across subtests for the older groups, and larger, across groups, for the Information and Vocabulary subtests.

The pattern of change for the Performance subtests, on the other hand, is inconsistent both across groups and across subtests. Nonetheless, there is a clearly distinguishable interaction between age and direction of change.

Means of absolute differences between first and second tests (in IQ points)*

	Younger		Older		Total SEM**	
	Short	Long	Short	Long		
Full-scale IQ	5	5	6	6	6	3
Verbal IQ	6	8	7	5	7	4
Performance IQ	7	7	6	8	7	5
Information	9	11	6	6	8	5
Similarities	8	9	10	9	9	5
Arithmetic	9	11	12	7	10	8
Vocabulary	10	12	7	9	10	4
Comprehension	10	6	7	11	9	7
Digit Span	12	14	9	9	10	9
MEDIAN	**9.5**	**11**	**8**	**9**	**9.5**	**6**
Picture Completion	8	7	9	9	9	6
Picture Arrangement	8	7	11	10	9	7
Block Design	10	8	9	6	9	5
Mazes	15	17	22	20	19	8
Coding	11	13	9	10	11	6
Object Assembly	13	12	9	11	12	7
MEDIAN	**10.5**	**10**	**9**	**10**	**10**	**7**

*One point on each subtest scale = 5 IQ points.
**Computed from the constancy coefficients in Table 11 of the manual for the Hebrew language version of the WISC-R (Lieblich, Ben-Shakhar-Segev, & Ninio,1976).

Table 2

Percentage of non-zero differences (test-retest) that are negative

	Younger		Older		Total
	Short	Long	Short	Long	
Full-scale IQ	57	64	79	78	70
Verbal IQ	70	72	75	83	75
Performance IQ	36	44	57	66	51
Verbal Subtests					
Information	89	88	65	80	80
Similarities	50	26	77	64	53
Arithmetic	58	78	53	56	63
Vocabulary	85	91	90	76	86
Comprehension	50	47	68	65	58
Digit Span	28	28	48	51	39
MEDIAN	**54**	**63**	**66**	**65**	**61**
Performance Subtests					
Picture Completion	54	36	61	70	54
Picture Arrangement	30	35	60	76	49
Block Design	26	41	44	42	39
Mazes	55	59	36	64	61
Coding	51	62	30	48	48
Object Assembly	63	35	50	67	52
MEDIAN	**53**	**39**	**47**	**66**	**51**

Table 3

Means of non-absolute test score differences (in IQ points)*

	Younger		Older		Total
	Short	Long	Short	Long	
Full-scale IQ	-1	-2	-4	-5	-3
Verbal IQ	-4	-5	-5	-5	-5
Performance IQ	2	2	-1	-4	0
Verbal Subtests					
Information	-5	-10	-3	-5	-6
Similarities	0	6	-7	-3	-1
Arithmetic	-3	-6	-2	-1	-3
Vocabulary	-9	-10	-6	-6	-8
Comprehension	2	1	-3	-4	-1
Digit Span	6	5	0	1	3
MEDIAN	**-1.5**	**-2.5**	**-3**	**-3.5**	**-2**
Performance Subtests					
Picture Completion	-1	2	-4	-5	-2
Picture Arrangement	5	3	-4	-4	0
Block Design	6	2	0	1	2
Mazes	3	6	0	8	-3
Coding	-2	-3	6	1	0
Object Assembly	-2	4	-1	-5	-1
MEDIAN	**1**	**2**	**-0.5**	**-1.5**	**-0.5**

*One point on each subtest scale = 5 IQ points.

Table 4

Retest scores on the Performance subtests (and the Performance IQ) were higher than scores on the initial test for younger children and lower for older ones. This is reflected both in a minority of negative differences (36% and 44% for the Short and Long time intervals, respectively) and in positive mean differences (2 IQ points for both time intervals). This age-by-direction interaction causes a larger decline in mean Full-scale IQ scores for children identified as gifted at an older age. It also explains the lack of change for the Performance IQ in the entire sample (a mean difference of 0 and an even proportion of positive and negative differences). This constancy in the Performance IQ is the main reason for the high percentage of subjects (86%, see Table 5) who were (again) defined as gifted at the second measurement.

Consistency of Change

In considering the consistency of score changes over subtests for each subject, we sought a pattern of some children who usually increased their scores and others who usually lowered them. Thus, we calculated for each research group the percentage of subjects whose Verbal and Performance IQ scores changed in opposite directions out of all subjects with score changes.

Percentages were relatively high, ranging from 34%-50%, suggesting that score changes were random and did not form a pattern. This finding is supported by evidence of the constancy of relative ranking within the group. Correlations between first and second test scores were relatively high for the Verbal, Performance and Full-scale IQ scores (.66, .61 and .64, respectively) and somewhat lower for the subtests.

DISCUSSION

The study elicited pleasantly surprising results: The IQ scores of children identified as gifted at an early age did not change considerably after an average time interval of 2.5 years. Mean absolute differences between scores at the two points of measurement ranged from 1/3 to 1/2 SD (5-8 IQ points) for the three IQ scores, and from 1/2 to 3/4 SD (8-11 IQ points) on the subtests, with no consistent differences attributable to age of identification.

It is difficult to determine how much of this difference is attributable to the retesting itself and how much to the time interval. There are, however, two indications that a considerable degree can be ascribed to the readministration of the same test. First, differences (albeit

Retest score distribution: "Gifted" vs "non-gifted" (N's and percentages)*

		Verbal IQ		
		106-129	130-153	
Performance IQ	93-129	22 (14%)	36 (22%)	58 (36%)
	130-155	37 (23%)	66 (41%)	103 (64%)
		59 (37%)	102 (63%)	161 (100%)

*Any subject whose Verbal or Performance score surpassed 129 IQ points was defined as gifted.

Table 5

of a smaller magnitude) were also found between scores obtained on near simultaneous administrations of the test in the standardization sample of the WISC-R (although the smallness of the sample on which the constancy coefficients are based cannot be ignored). A second indication is our own finding that the time interval between tests (1-4 years) had no consistent effect on mean absolute differences between initial test and retest scores. This suggests that the time interval plays a rather small role relative to the retesting. Of course, four years is a relatively short span of time; the effect of the time span is likely to be more salient at much greater intervals.

A clear distinction was found between Performance and Verbal scores. Whereas the change in the former was inconsistent between testings, the latter clearly declined at the second measurement (by 4-6 IQ points) in all four groups. This downward trend is largely attributable to score differences in two subtests: Information and Vocabulary. Lincoln (1935) had similar findings of a decline in Verbal scores, particularly for Abstract Words and Vocabulary. His study, an extension of the Harvard Growth Study, one of the largest longitudinal studies conducted in the United States (Dearborn & Rothney, 1941), is similar to the present one in terms of the initial age of testing and time interval between measurements, but differs from it in terms of the test administered to the children (Stanford-Binet).

Results did not point to intra-individual consistency between changes in the Verbal and Performance sections of the test. Indeed, for half of the subjects, Verbal and Performance IQ scores changed in opposite directions.

The decline in the IQ scores of the gifted children at the second measurement is to be expected, as this group was selected on the basis of its extremely high scores on the first test. Given the imperfect correlation between any two measurements for the general population, the mean scores of this highly selective population could only decrease. Indeed, the decline was greater on those tests for which mean scores at the first measurement were particularly high (Verbal IQ), and it was negligible when initial means were relatively low (Performance IQ). What is somewhat surprising is the finding that mean scores sometimes remained constant or even rose (particularly the Mazes subtest). These results may be at least partially attributable to random differences between age groups in the standardization sample of the WISC-R test.

These random differences may also explain the unexpected finding of a greater decline in the Full-scale IQ score for children identified as gifted at a later age. It should be noted that Hildreth (1943) had similar results. Employing a similar sample size, measurement interval and criterion for defining gifted children (but using the Stanford Binet Test), Hildreth found the percentage of negative differences among the group identified at a later age (over 10) to be relatively larger than that of the group identified at an earlier age (below 10). In contrast, Terman (1930) found no connection between the tendency to decline at the second testing and age of identification.

As opposed to Lincoln's (1935) findings that IQ scores declined more when the period between testings was longer, the present study found no clear effect of measurement intervals on test score changes. The lack of an effect may be due to the smallness of the overall time interval (4 years).

The extent of IQ score decline revealed in the present study was less than that generally found in the literature. Indeed, 86% of the subjects were again defined as gifted on the basis of the second test administration. These results support the assumption of "constancy" that underlies early identification of students for gifted programs on the basis of IQ test scores, at least throughout the first years of elementary school. It should be stressed, however, that they should not be viewed as a wholesale justification for the practice of identifying gifted children only once. After all, the time interval covered by the study was relatively short. More importantly, this interval did not include the adolescent years, when major changes are expected.

REFERENCES

Cattell, P (1933) Do the Stanford-Binet IQ's of superior boys and girls tend to decrease or increase with age? *Journal of Educational Research, 26,* 668-673

Dearborn, W F & Rothney, J (1941) *Predicting the child's development* Cambridge, Mass Science-Art

Feldhusen, J F , & Hoover, S M (1986) A conception of giftedness intelligence, self concept and motivation *Roeper Review, 8,* 140-143

Hildreth, G (1943) Stanford Binet retests of gifted children *Journal of Educational Research 37,* 297-302

Hollingworth, L S (1938) An enrichment curriculum for rapid learners at public school 500 Speyer School *Teachers College Record, 39,* 296-306

Hollingworth, L S , & Kaunitz, R M (1934) The centile status of gifted children at maturity *Journal of Genetic Psychology, 45,* 106-120

Lieblich, A , Ben Shakhar-Segev, N , and Nimio, A (1976) *WISC-R Guide to the Wechsler Test for Israeli Children* Jerusalem The Institute for Developmental Psychology, The Hebrew University of Jerusalem and the Psychological Consultation Services, The Ministry of Education and Culture (Hebrew)

Lincoln, E A (1935) A study of the changes in the intelligence quotients of superior children *Journal of Educational Research, 29,* 272-275

Matarazzo, J D (1972) *Wechsler's measurement and appraisal of adult intelligence* (5th ed) Baltimore, MD Williams & Wilkins

Silverman, L K (1986) The IQ controversy - conceptions and misconceptions *Roeper Review, 8,* 136-140

Sternberg, R J (1986) Identifying the gifted through IQ Why a little bit of knowledge is a dangerous thing *Roeper Review, 8,* 143-147

Tannenbaum, A J (1983) *Gifted children* New York Macmillan Terman, L M (1925) *Genetic studies of genius, Vol II Mental and physical traits of a thousand gifted children* Stanford, Calif Stanford University Press

Terman, L M (1930) *Genetic studies of genius, Vol III The promise of youth* Stanford, Calif Stanford University Press

Thorndike, R L (1940) Constancy of the IQ *Psychological Bulletin, 37,* 167-186

Thurstone, L L (1926) The mental age concept *Psychological Review, 33,* 4, 268-278

Treffinger, D J , & Renzalli, J S (1986) Giftedness as potential for creative productivity Transcribing IQ scores *Roeper Review, 8,* 136-140

Wechsler, D (1974) *Manual for the Wechsler Intelligence Scale for Children* - Revised New York The Psychological Corporation

Witty, P A (1940) A genetic study of fifty gifted children *Thirty Ninth Yearbook of the National Society for the Study of Education* Part II, 401-409

Corrections

The publication information for the book, *Understanding Those Who Create,* by Jane Piirto, Vol.15, no.2, page 110, should read:
Ohio Psychology Press, 1992 360 pp. $20.00