Stability of WISC-R Scores between Triennial Evaluations of Learning Disabled Students

Norman J. Wozny
University of Nebraska at Omaha

Follow this and additional works at: https://digitalcommons.unomaha.edu/studentwork

Part of the Psychology Commons, and the Special Education and Teaching Commons

Recommended Citation
https://digitalcommons.unomaha.edu/studentwork/142
Stability of WISC-R Scores between Triennial Evaluations of Learning Disabled Students

A Field Project
Presented to the
Department of Psychology
and the
Faculty of the Graduate College
University of Nebraska at Omaha

by
Norman J. Wozny
June 4, 1992
Field Project Acceptance

Accepted for the faculty of the Graduate College, University of Nebraska at Omaha, in partial fulfillment of the requirements for the degree, Specialist in Education, University of Nebraska at Omaha.

<table>
<thead>
<tr>
<th>Name</th>
<th>Department</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deana T. Lidd</td>
<td>Psychology</td>
</tr>
<tr>
<td>Jon McCall</td>
<td>Special Edu.</td>
</tr>
<tr>
<td>Norman H. Jones</td>
<td>Psychology</td>
</tr>
</tbody>
</table>

Chairperson: Deana T. Liddy
Date: June 4, 1992
Acknowledgments

I would like to express appreciation to my committee chair, Dr. Deana Liddy, and members of my committee, Dr. Norman Hamm and Dr. Tom Lorsbach, for their patience and guidance throughout this project. I would like to extend special thanks to Dr. Richard Wikoff for his assistance with the analysis of variance portion and to Dr. David Probst of Fremont Public Schools for his cooperation and assistance with my gathering of data in his district. Finally, I would like to express appreciation to Charlene, my wife, for her assistance in recording data for this project and for her patience and understanding throughout this project as well as an internship and a number of other major undertakings.
# Table of Contents

Page

Abstract .................................................................................. vi

Introduction ................................................................................. 1

  LD Defined .............................................................................. 2

  Implications of Score Stability ............................................... 5

  Review of the Literature ......................................................... 5

  Statement of the Problem ....................................................... 18

Method .................................................................................. 18

  Subjects ................................................................................ 18

  Statistical Analyses ................................................................ 22

Results .................................................................................. 23

  ANOVA ................................................................................. 23

  Pearson r ............................................................................. 26

  Changes in Individual Scores ............................................... 26

  Regression to the Mean ......................................................... 29

Discussion ............................................................................. 30

  Statistical Analyses .............................................................. 30

  Regression to the Mean ........................................................ 34

  Limitations ............................................................................. 34

  Suggestions for Research ........................................................ 35

  Educational Implications ....................................................... 36
List of Tables

Table 1. Highlights of Studies Examining WISC-R Score Stability between Reevaluations of Learning Disabled Students

Table 2. Comparison of Percentages of Significant Shifts in IQ Scores in Studies Presenting Frequency Data

Table 3. Means and Standard Deviations of First and Second Administrations of WISC-R VIQ, PIQ and FSIQ

Table 4. Repeated Measures Analysis of Variance Table for First and Second Administrations of VIQ by Age Group

Table 5. Repeated Measures Analysis of Variance Table for First and Second Administrations of PIQ by Age Group

Table 6. Repeated Measures Analysis of Variance Table for First and Second Administrations of FSIQ by Age Group

Table 7. Frequency of Significant Shifts in VIQ, PIQ and FSIQ Scores between Administrations Based on Bands of Error

Table 8. Frequency of Significant Shifts in VIQ Scores by Age Group Based on Bands of Error

Table 9. Frequency of Significant Shifts in PIQ Scores by Age Group Based on Bands of Error

Table 10. Comparison of Mean Scores from Time 1 to Time 2 of Individuals at the Upper and Lower Extremes
Abstract

Recent studies of intelligence test score stability among learning disabled children have reported adequate stability when correlational and analysis of variance (ANOVA) techniques were used. However, less than adequate score stability has been found when individual scores were examined.

The present study explores the test-retest stability of the Wechsler Intelligence Scale for Children (WISC-R) using three statistical methods: Pearson product-moment correlation, analysis of variance, and an examination of individual scores. Regression to the mean is also examined.

While reasonably high levels of stability are concluded by the Pearson product-moment correlations, significant drops in Verbal and Full Scale IQ scores between administrations are revealed by the ANOVA and by the examination of individual scores in the 9, 10 year old age group.

The discussion includes implications for educational programming and for methods of evaluation, including alternative approaches to reevaluations. Suggestions for further research are also presented.
Introduction

Stability of intelligence test scores is an issue that has implications for the identification and reevaluation of learning disabled (LD) students. LD students represent a significant portion of the population and comprise the largest and fastest growing handicapped group in education (Lerner, 1988).

A number of authors have pointed out that most studies of intelligence test score stability have focused on the normal population, with only a few studies investigating the LD population (Anderson et al, 1989; Oakman et al, 1988; Schmidt et al, 1989; Stavrou, 1990; Webster, 1988).

The amount of expected change in intelligence test scores over time has been studied extensively and is predictable within the normal population (Stavrou, 1990; Webster, 1989). However, these data cannot necessarily be generalized to special populations. More information about intelligence test score stability is needed for the learning disabled population because of implications regarding verification of handicapping conditions, special education placement, instructional programming and even the appropriateness of the use of intelligence testing in reevaluations (Anderson et al, 1989; Oakman et al, 1988; Stavrou, 1990).

Of the existing studies of intelligence test score stability among learning disabled students, many have been
criticized for various methodological problems including too few subjects, too short an interval between test administrations and for overlooking important facts by limiting analyses to group data (Anderson et al, 1989; Oakman et al, 1988; Stavrou, 1990; Vance et al, 1987; Webster, 1988).

Lerner (1988) presented data on the estimated incidence of learning disabilities. Students with learning disabilities represent the largest handicapped group defined by PL 94-142. Estimates of the percentage of the general student population in the United States who have learning disabilities range from 1 percent to 15 percent among various researchers. In 1968, the National Advisory Committee on Handicapped Children recommended to Congress a conservative estimate of 1 to 3 percent pending more objective evidence. The actual count of students, ranging in age from 3 to 21 years, receiving special services under the classification of learning disabled is 4.8 percent of the student population (figures are for school year 1986-87). Figures spanning the years from 1977 to 1987 indicate an increase in the learning disabled population from 1.89 percent to 4.8 percent of the general student population (Lerner 1988).

LD Defined

Lerner (1988) outlines various definitions of "learning disability" which have been proposed over the years by
governmental and private groups who represent the best interests of children and adults with this impairment to learning. Of these, the most widely used is the federal definition incorporated in PL 94-142, the Education for all Handicapped Children Act of 1975.

One major concept employed in the federal definition is a severe discrepancy between a student's apparent potential for learning and his/her level of achievement. Many states, including Nebraska, have adopted a discrepancy formula to identify students who qualify for special education services under the classification of learning disabled. Lerner cites a study by Frankenberger and Harper (1987) which found that by 1986, 57 percent of the states used a discrepancy formula in identifying learning disabilities.

States and school districts determine what constitutes a severe discrepancy. Intelligence tests are generally used to determine potential for learning. Under the discrepancy formula, achievement scores in one or more areas are compared to an intelligence test score to determine whether a difference exists, between a student's potential for learning and his/her achievement, sufficient to constitute a learning disability. For verification of a learning disability, Nebraska law (Rule 51) requires that achievement levels in one or more of seven academic areas (basic math, applied math, basic reading, reading comprehension, oral expression, written expression and listening comprehension)
must appear at least 1.3 standard deviations (SD), equating to 20 standard score points, below the intelligence test score of the individual in question. The composite intelligence test score is used as a reference unless a discrepancy of 1 S.D. or greater exists between the component intelligence test scores. In that case, the higher of the two component intelligence test scores may be used as a reference for comparison in the discrepancy formula.

The Wechsler Intelligence Scale for Children-Revised (WISC-R) is commonly used as the measure of intelligence in the verification of learning disabilities. The Full Scale Intelligence Quotient (FSIQ) would typically be used as the reference for learning potential unless a 15 point or greater differential exists between the Verbal Intelligence Quotient (VIQ) and the Performance Intelligence Quotient (PIQ). In that case the higher of the latter two scores would be used. Standardized achievement test scores are compared to the WISC-R Full Scale or component IQ score in order to determine the amount of the discrepancy. In addition to the required 1.3 S.D. discrepancy, Nebraska Rule 51 stipulates that, in order to verify a learning disability, the achievement score must be at or below the 25th percentile, or in terms of a standard score, 90 or below. The IQ score must be at or above 1 SD below the mean, equating to a score of 86 or greater on the WISC-R.
Implications of Score Stability

The implications surrounding the issue of stability of WISC-R scores are apparent when one considers the discrepancy formula. A student who is verified as learning disabled by the discrepancy formula may or may not be reverified on a subsequent mandatory evaluation three years later, or may be identified as having a different handicapping condition such as Mentally Handicapped - Mild.

Further implications were presented by Stavrou (1990) who notes that special education placements are based on the assumption of stable intelligence test (IQ) scores. She suggested that an increase in IQ score over time may imply that identification by IQ may not be appropriate, and that a decrease in IQ score may cast doubt on the efficacy of special education placement programs.

Stavrou further stated that the constancy of IQ scores is well documented in the general population, but that few studies have examined IQ score stability in special populations, especially the LD population. Stavrou pointed out that information regarding the stability of IQ scores among LD and mildly retarded children is of particular importance to the school psychologist.

Review of the Literature

Stavrou has determined that the various methods used in measuring IQ score stability may yield differing results. She studied the longitudinal stability of WISC-R scores in
100 LD and 60 mildly retarded children tested on three different occasions. The mean age at the first testing of the LD group was 8-8, the MMR mean age was 8-11; the second testing LD mean age was 12-2 and the MMR mean age was 13-10; the third testing LD mean age was 14-10 and the second testing MMR mean age was 15-6. Stavrou reported fairly stable FSIQ scores and less stable VIQ and PIQ scores over a period of six years for both samples, using analysis of variance (ANOVA) and correlational methods. Greater variability was observed in both samples when the frequency of significant changes in individual scores was examined.

Stavrou acknowledged that this study is limited in its generalizability because students selected as subjects were limited to those in suburban Long Island, NY schools. A selection bias may have been present in that this group may have been representative of relatively stable residents in this six year study. Other, more transient students were not included because data for the six year period was not available for them.

Webster (1988) stated that few studies exist on the temporal stability of IQ scores. He criticized those few for focusing on the normal population, for which the annual decline in test performance is predictable, and for not examining the effects of a cognitive disorder on temporal stability. Additional criticism levied by Webster cited too few subjects, pooling subjects into too wide an age range,
and using group reliability coefficients rather than examining the number of individuals with scores that changed substantially between administrations.

Webster studied 83 LD students and 72 EMR students between the ages of 13-6 and 14-11 (age at first administration) in rural North Carolina schools. LD was defined as a discrepancy of two or more years between expected grade level and actual achievement, with expected grade level determined by the subjects' WISC-R scores. The diagnosis of EMR was assigned to subjects whose IQ scores fell between 50 and 69, +/- 1 SD. Subjects were administered the WISC-R on two occasions for the purpose of mandatory three year evaluations of special education placements. The mean test-retest interval was three years, one month.

Webster analyzed the data by using a three step procedure. First, group reliability coefficients were computed by using Pearson r correlations. Reliability coefficients in the range of .90 were found for the LD group and in the .50 and .60 range for the EMR group. Greater stability of IQ scores was found for the LD group than for the EMR group.

Second, a repeated measures ANOVA was applied to the group means of the VIQ, PIQ and FSIQ scores with no significant differences found between the first and second administrations for either group.
Finally, each case was analyzed for shifts in VIQ, PIQ and FSIQ. 22.9% of the LD group shifted in a positive or negative direction on both the VIQ and the PIQ by an amount equal to or exceeding the standard error of measure ($SE_M$) at the 95% level of confidence. Significant shifts were demonstrated by 22% of the EMR group in the VIQ score, and by 28% in the PIQ score. 28.9% of the LD group and 15% of the EMR group demonstrated significant shifts in the FSIQ score.

Webster found the scores of the LD group more stable than those of the EMR group. He concluded that age and cognitive development are factors that impact on IQ score stability. Webster selected adolescents for his study in order to avoid introducing a possible confounding variable of a transitional learning style which might be present in children who have not yet attained formal operations. This appears to be an untested presumption by Webster.

A number of other recent studies have also examined IQ test score stability among exceptional students, particularly LD. Most have used a substantial test-retest interval, usually geared to mandatory three year reevaluations. Results of these studies have varied in terms of their findings regarding stability of VIQ, PIQ and FSIQ scores and in interpretations of results, especially implications regarding the necessity for three year reevaluations.
Oakman and Wilson (1988) examined 150 LD students, ranging in age from 9.5 to 16.5 years, who had been readministered the WISC-R for triennial reevaluations. Three statistical methods were utilized in analyzing the data. A correlated t-test was used to test differences between the means of the VIQ, PIQ and FSIQ scores; ANOVA to test for possible interaction between score stability and severity of learning disability; and a frequency distribution of individual changes in FSIQ scores.

Significant differences were found between mean scores on the PIQ and FSIQ scales. No significant differences in score stability were found among the different levels of severity of learning disability, as defined by type of special education program placement.

A frequency distribution of FSIQ score differences revealed changes of 4 points or less for 52% of the sample and changes of 15 points or more for only 4% of the sample. However, the authors failed to point out that the remaining 44% of the sample demonstrated changes of 5 to 14 points. No figures were presented for this portion of the sample which may have contained a substantial number of significant changes in FSIQ scores. No frequency distribution was presented for the VIQ and PIQ scales which may have revealed important shifts in scores and may possibly have had a compensatory effect and impacted the incidence of changes in the FSIQ scores.
The authors concluded that WISC-R scores are sufficiently stable to eliminate the need for readministration of this instrument for triennial evaluations. However, the methodological problems discussed here suggest that this conclusion was premature and that a more complete analysis of this data is needed.

A six year longitudinal study by Vance, Hankins and Brown (1987) examined WISC-R stability in a sample of 20 learning disabled and 12 mentally handicapped students ranging in age from 6-5 at first testing to 16-11 at the third testing. Data were analyzed using repeated measures ANOVA and product-moment correlations.

The authors concluded that reliability coefficients were unsatisfactory, having established .80 as satisfactory, based on a statement to that effect by Sattler (1982). The repeated measures ANOVA found no significant changes in IQ scores across the three testing periods, although the authors stated in passing that large changes in individuals' scores were noted. However, no data were presented regarding the incidence of changes in the scores of individual subjects.

The undifferentiation of the learning disabled and the mentally handicapped subjects in this study has introduced a methodological problem in terms of generalizability. Other studies have found significant differences in IQ score stability between those two groups (Stavrou, 1990; Webster,
Therefore, the data cannot be generalized to either group.

Schmidt, Kuryliw, Saklofske and Yackulic (1989) studied WISC-R score stability in a sample of 36 LD students with a mean test-retest interval of 2.5 years. Mean ages of the subjects at the times of the initial and follow-up testing were 8.1 years and 10.1 years respectively. A correlated \( t \)-test was used to compare the means of the first and second test administrations. Significant decreases were reported in VIQ (\( p < .01 \)) and FSIQ (\( p < .005 \)) between the first and second administrations. No significant change was reported in PIQ.

Incidence of change in individual test scores by one SEM or greater was also reported. One SEM for the VIQ, PIQ and FSIQ scales equals +/-7, +/-9 and +/-6 respectively. Decreases of one SEM or greater were reported for 42% of the subjects on VIQ, 14% on PIQ and 39% on FSIQ. Increases of one SEM or greater were reported for 11% of the subjects on VIQ, 8% on PIQ and 14% on FSIQ.

The authors attributed the decreases in FSIQ scores to the decreases in VIQ scores. It was further suggested that the decreases in VIQ may be due to a secondary effect among children with reading decoding problems, manifested in a failure to acquire information and vocabulary concepts which are normally acquired through reading.
The main focus of this study was on identifying subtypes of learning disabilities and no conclusions were drawn as to the use of the WISC-R in reevaluations except to point out the authors' perspective on its limitations as a diagnostic instrument. However, the results suggested that a substantial number of individuals demonstrated significant changes in IQ scores.

Anderson, Cronin and Kazmierski (1989) studied WISC-R score stability in 113 LD students over a period of three years. Mean ages for first and second administrations were 8 years, 3 months and 11 years, 7 months respectively.

A t-test analysis found a significant difference between times of testing in the VIQ score only, and represented a decrease by an average of 4.1 points.

Pearson r correlations were .55, .63 and .58 for the VIQ, PIQ and FSIQ scores, respectively.

An examination of individual scores revealed that 64% of the sample experienced a decrease in VIQ, 32% showed an increase and 4% remained the same. No further information is reported on the incidence of change in VIQ scores and none is reported for the PIQ and FSIQ scores.

The authors concluded that this study found much lower reliability than other studies of special populations have found, although the phenomena of decreased VIQ and increased PIQ scores has been documented previously. The authors suggested that these changes may possibly be explained by
advancing difficulty with verbal conceptualization and abstract verbal thinking experienced by handicapped students as grade level advances, or that special education classes may stress development of perceptual-motor skills at the expense of verbal instruction. In terms of implications for practice, the authors concluded that because of insufficient IQ test score stability, it is probably advisable to readminister the WISC-R on triennial evaluations.

Schuerger and Witt (1989) conducted a meta-analysis of 34 studies which examined factors affecting the temporal stability of intelligence test scores. Data were included on the test-retest reliability of five intelligence scales, the Stanford-Binet Intelligence Scale: Fourth Edition, the Wechsler Intelligence Scale for Children (WISC), the WISC-R, the Wechsler Adult Intelligence Scale (WAIS) and the Wechsler Adult Intelligence Scale - Revised (WAIS-R).

Multiple regression procedures were used to find predictors of test-retest reliability of IQ scores, the dependent variable. The independent variables included age (ranging from 3 to 65), status (two levels called normal and patient, the former referring to subjects of normal or low IQ including LD and mentally handicapped, and the latter referring to patients hospitalized for mental or physical disorders), gender, and test-retest interval (ranging from .25 months to more than 280 months).
Only two factors, the age at first testing and the test-retest interval, were found to be significant predictors of test-retest reliability. These two predictors accounted for more than 50 percent of the variance in reliability. As age at first testing increased, reliability coefficients increased, and as test-retest intervals increased, reliability coefficients decreased. These two variables were also applied to the number of subjects experiencing changes in IQ score of 15 points or more (1 SD) with similar results.

The Schuerger and Witt study is presented here as a point of reference to factors affecting IQ score stability in the general population. Subjects were not grouped by cognitive status and included normal IQ levels as well as low IQ levels which included LD and mentally handicapped subjects.

Studies of intelligence test score stability among the normal population have documented the effect of the length of the test-retest interval on score stability. The stability of intelligence test scores has been shown to decline as the test-retest interval increases (Bayley, 1949). Webster (1988), citing Eysenck (1953), points out that temporal stability estimates (using correlational procedures) of intelligence test scores of normal subjects have even been shown to decline by an annual rate of .04.
The studies reviewed herein have found varying results in terms of the sufficiency of intelligence test score stability. However, those which present data on the incidence of significant changes in test scores over time have reported frequency distributions which indicate sufficient numbers of substantial changes in IQ scores to cast doubt on the findings of high test-retest stability (Schmidt et al, 1989; Stavrou, 1990; Webster, 1988). A summary of studies is presented in Table 1.

Insert Table 1 about here

Three studies reviewed herein have presented data on significant changes in individual scores. All three consistently reported high percentages of subjects with substantial changes in scores over a 3 year period. The three studies have used somewhat different criteria for significance, but are roughly comparable at the 95% level of confidence. They have reported significant changes in 18% to 53% of the subjects examined. As a criterion for significant change, Schmidt et al (1989) adopted a band of error presented by Kaufman (1979). Webster (1988) used a formula which doubled the $SE_M$ from the WISC-R standardization data, and Stavrou (1990) used a $SE_M$ of difference, a procedure for evaluating significant changes
in individual scores. A comparison of percentages of significant changes is presented in Table 2.

In analyzing the data, most of the studies reviewed herein have used Pearson r correlations and ANOVA. Because both of these methods are averaging processes, some important differences are masked. It is possible that the compensatory effects inherent in these techniques has not revealed potentially high incidences of significant shifts in individual IQ scores over time.

Researchers of intelligence test score stability related to triennial evaluations have generally not commented on achievement test scores as a further source of variability in discrepancy formula applications. It becomes apparent that, with questionable WISC-R score stability and with achievement tests presenting an additional source of potentially unstable scores, the appropriateness of the use of the discrepancy formula becomes an area of concern.

Ferguson and Mamen (1985), in an article calling for more comprehensive evaluations of LD students, have cited some unsatisfactory psychometric characteristics in individual achievement tests. The authors have pointed out that, while a child may place above or below average in one classroom, he/she may place differently in another classroom
and that compelling arguments for locally standardized tests have been presented.

A second criticism presented by the authors was that, in some situations, a child may be compared to children of a younger age on an achievement test. This might occur in the case of a child who has repeated one or more grades.

The authors have nonetheless recognized that achievement testing, however it is accomplished, is important in the identification of learning disabilities. Furthermore, they list some specific basic areas of achievement that should be assessed.

Varying conclusions among researchers regarding the adequacy of the test-retest stability of the WISC-R have pointed to differences in their perceptions of the usefulness of that instrument. Oakman and Wilson (1988) have suggested that no new information is gained by readministering the WISC-R. Conversely, Vance et al (1981) have encouraged readministration for the purpose of defining specific skills and using the information obtained for educational programming. Schmidt et al (1989) have attributed only limited diagnostic capability to the WISC-R in defining strengths and weaknesses in exceptional populations and have concluded that it is only useful as a measure of global intelligence. They have suggested a need for further research in the area of defining categories of learning disabilities by using more specialized measures.
Statement of the Problem

The impact of the stability of WISC-R scores on the verification of a learning disability by application of a discrepancy formula has been discussed previously and is a major focus of the present study. The inconclusive results reported in the literature point to a need for more knowledge of WISC-R score stability in the learning disabled population. Research on the effects of age differences on score stability is lacking and more data are needed from diverse geographic regions.

Method

In the present study, the test-retest stability of WISC-R IQ scores over a three year period was investigated for a population of students who have been verified as LD in accordance with Nebraska Rule 51. Test-retest stability was examined using correlational and analysis of variance techniques. Age effects were examined using analysis of variance. Changes in individuals' scores between test administrations were examined, and an examination of regression to the mean was conducted.

Subjects

The subjects in this study were 81 students who were verified LD by a discrepancy formula in accordance with Nebraska Department of Education requirements. Subjects were drawn from the files of an educational service unit serving public schools in a four county rural area in
Northeast Nebraska, and from a school district in a small city (population 25,000) in the same geographic area.

Subjects were selected on the basis of several criteria. First, availability of scores for two separate testings over a period of approximately three years was necessary for each subject. Thus, at least two WISC-R protocols and associated achievement test scores for each evaluation were necessary for selection. Strict adherence to the discrepancy formula was a condition of selection. That is, no subjects were selected whose records indicated any deviation from the twenty point discrepancy or whose protocols indicated any omissions of subtests, substitutions of subtests or prorations of WISC-R IQ scores. In this way a consistent operational definition of LD was established in case other criteria, such as clinical judgment or past discrepancy requirements, which may differ with current ones, were used by multidisciplinary teams for identifying LD students in earlier years. Subjects verified with a dual handicapping condition of learning disabled/behavior disordered (LD/BD) were excluded in order to control for possible effects of BD.

The result of the most recent evaluation was not a factor in the selection of subjects. That is, whether a subject was reverified as LD, was verified under a different handicapping condition, (e.g. Mentally Handicapped - Mild), or was not verified with a handicapping condition was not
considered in selecting subjects, although these data were recorded. In a study of temporal stability of WISC-R scores, Webster (1988) excluded the data of subjects who were not reverified as LD on their most recent evaluations. This strategy was an attempt to control for a change in the discrepancy formula which occurred during the test-retest interval. It is possible that important data may have been lost by imposing LD criteria at the end of the test-retest interval. No such criteria were imposed by the present study in which the focus was to examine changes in WISC-R scores of students who were verified LD during the test-retest interval, no matter what those changes were. To control for any possible changes in the LD formula, the current discrepancy criteria were imposed, as discussed above. To eliminate scores obtained at the end of the term is to potentially throw out changes which may be program related.

Only 20 percent of potential subjects (i.e. those with LD verifications) were selected for the current study. Obviously, many more subjects were rejected than were accepted for inclusion in the study. Reasons for the exclusion of subjects included a number of conditions relating to the available data. Examples included the following: multidisciplinary team verification without a twenty point discrepancy, WISC-R subtests substituted or omitted, IQ scores prorated, use of intelligence scales
other than the WISC-R, not all data present (e.g. achievement test scores), and only one evaluation on file. In some cases the Mazes subtest was substituted for the Coding subtest by the examiner. When such instances appeared during the present data compilation, the scores were recomputed using the appropriate subtest and, if the discrepancy was valid, the data were included.

The total number of 81 subjects included 58 males and 23 females. Grades range from Kindergarten to grade 8 for the first administration and from grade 3 to grade 11 for the second administration.

The mean age of the subjects at the first administration was 10 years, 6 months, with a minimum age of 6 years, 7 months and a maximum age of 14 years, 1 month. The mean age at the second administration was 13 years, 6 months, ranging from 9 years, 2 months to 16 years, 11 months.


The mean test-retest interval was 36.11 months with a minimum interval of 28 months and a maximum interval of 45 months.

Of the 81 subjects, 59 were reverified LD on the second administration of the WISC-R in accordance with the
discrepancy formula of the Nebraska requirements. 15 were reverified LD but did not fit the discrepancy formula. One subject was verified Mentally Handicapped-Mild at the second administration and 6 were verified with no handicapping condition. Therefore, of this sample of 81 LD students, 59 fit the discrepancy formula on reevaluation while 22 no longer met the discrepancy requirements for identification as LD.

Statistical Analyses

Three different statistics were used in analyzing the data.

A 4 (age group) x 2 (time) repeated measures ANOVA was performed for each of the dependent measures, VIQ, PIQ, and FSIQ to test for significant shifts in scores between the first and second administrations, to test for effects of four age groups at the time of the second administration and to test for the interaction of age group and time of administration.

A within subjects one-way ANOVA over times of administration was performed for the VIQ and FSIQ to test for significant shifts in score within each age group.

Pearson r correlations were calculated for the test-retest interval of the VIQ, PIQ and FSIQ in order to examine changes in the relative positions of the subjects between the first and second administrations.
Shifts in individual VIQ, PIQ and FSIQ scores were examined by a frequency distribution enumerating the absolute value of the shifts in scores between administrations which are outside the band of error at the 95% confidence level and those at the 99% confidence level (Kaufman, 1979). The percentage of N represented by those absolute values is also reported.

Finally, extreme scores were examined for regression to the mean.

Results

The means and standard deviations for each age group and for the total sample for both administrations of VIQ, PIQ and FSIQ are presented in Table 3.

Analysis of Variance

VIQ

A 4 (age group) x 2 (time of administration) repeated measures ANOVA was performed for VIQ. The four age groups were designated as follows: 9, 10; 11, 12; 13, 14; 15, 16; and represent subjects' ages at the time of the second administration. The time of the first administration was designated VIQ1 and the time of the second administration was designated VIQ2. The ANOVA summary table for VIQ is presented in Table 4.
Significant main effects were found for both time of administration, $F(1, 81) = 13.07, p < .001$, and for age group $F(3, 80) = 4.06, p < .01$. The grand means for the first and second administrations of VIQ equal 93.004 and 90.454 respectively. VIQ grand means of each age group across times of administration follow: $9, 10 = 94.709; 11, 12 = 96.0; 13, 14 = 90.021; 15, 16 = 86.185$. Generally, the two older groups scored lower than the two younger groups.

Significant interaction effects of age group and time of administration were also found, $F(3, 81) = 7.22, p < .001$. In order to isolate the effects of the interaction of age group and time of administration, a within subjects one-way ANOVA was performed for each of the four age groups. Significant effects for time of administration were found in two age groups, $9, 10, F(1, 11) = 14.71, p < .003$, and $15, 16, F(1, 26) = 6.78 p < .015$. Table 3 presents the means. Means decreased for the youngest and oldest age groups and showed no significant change for the middle two age groups.

---

Insert Table 4 about here

---

**PIQ**

A repeated measures ANOVA found significant main effects for time of administration, $F(1, 81) = 6.55, p < .012$, and for age groups, $F(3, 81) = 2.72, p < .050$. The grand means for the first and second administrations of PIQ
equal 104.990 and 102.514 respectively. PIQ grand means of each age group across times of administration follow: 9, 10 = 100.833; 11, 12 = 107.772; 13, 14 = 99.896; 15, 16 = 106.556.

No significant effects of interaction of age group and time of administration were found, which means that there was no significant difference in the change in PIQ scores between any of the age groups. Therefore, no further ANOVA tests were applied to PIQ. The ANOVA summary table for PIQ is presented in Table 5. Table 3 presents the means.

A repeated measures ANOVA found significant main effects for time of administration, $F(1, 81) = 17.03, p < .001$. The grand means for the first and second administrations of FSIQ equal 98.255 and 95.488 respectively. Significant effects of the interaction of time of administration and age group were also found, $F(3, 81) = 5.12, p < .003$. No significant main effects were found for age group. The ANOVA summary table for FSIQ is presented in Table 6.

Since the interaction effect was significant, a within subjects one-way ANOVA for each age group was used to further isolate the effects of interaction. Table 3
presents the means. The one-way ANOVA found significant effects for time of administration for two age groups, 9, 10, \( F(1, 11) = 15.85, p < .002 \), and 15, 16, \( F(1,26) = 9.22, p < .005 \). The greatest decline in FSIQ was noted in the 9, 10 age group, 7.75 points. The 15, 16 age group experienced a decline of 2.48 points and the other groups experienced nonsignificant changes.

---

**Pearson r Correlation**

The Pearson r correlation was employed in further examination of the data. The correlation coefficients for each score across time are as follows: \( r(VIQ) = .8235 \), \( r(PIQ) = .7588 \), \( r(FSIQ) = .8205, p < .01 \). An adequate level of test-retest stability is noted for VIQ and FSIQ, with PIQ showing less stability. Sattler (1982) states that reliability coefficients of .80 or higher are generally considered acceptable for most cognitive tests.

**Changes in Individual Scores**

In order to examine shifts in individual IQ scores, a frequency distribution of changes in VIQ, PIQ and FSIQ was constructed for the entire sample. To further examine age effects, a frequency distribution was constructed for each of the four age groups within each of the two component scores, VIQ and PIQ, and for the FSIQ score. The 95% and
99% levels of confidence of bands of error proposed by Kaufman (1979) were applied to the examination of individual scores. Bands of error for VIQ, PIQ and FSIQ at the 95% level of confidence are +/- 7, +/- 9 and +/- 6 respectively. Bands of error for VIQ, PIQ and FSIQ at the 99% level of confidence are +/- 9, +/- 12 and +/- 8 respectively.

An examination of the incidence of individual changes within the entire sample of 81 students revealed a substantial number of shifts in VIQ, PIQ and FSIQ scores which fall outside the band of error at both the 95% and 99% levels of confidence. The number of individual scores and the percentage of N falling outside the band of error is presented in table 7. A somewhat higher number of significant shifts are noted for the VIQ than the PIQ. At the 99% level of confidence a significantly higher ratio of decreases to increases is noted for the VIQ compared to PIQ.

Insert Table 7 about here

An examination of the frequency distributions of the VIQ score shifts for each of the four age groups revealed large differences in percentage of N, with score shifts outside the band of error, between the age groups. Differences in ranges of change are also noted between age groups. The percentage of N and number of significant VIQ
score shifts are presented in Table 8 in terms of numbers of significant changes and percentage of N.

____________________________
Insert Table 8 about here
____________________________

The frequencies of shifts in PIQ scores and percentage of N undergoing shifts outside each band of error for the four age groups are presented in Table 9. The lowest rate of significant change was seen in the 15, 16 age group, while the highest rate was seen in the 13, 14 age group. The numbers of significant shifts in PIQ scores were less diverse between the four age groups than in VIQ shifts.

____________________________
Insert Table 9 about here
____________________________

The frequencies of shifts in FSIQ scores and percentages of N undergoing shifts outside the band of error are presented in Table 10. The highest rate of significant shift was seen in the 9, 10 age group and was similar to the degree of shift noted for VIQ in the same group. Therefore, the shifts seen in the FSIQ may be a reflection of VIQ score shifts. No significant upward score shifts were seen in the 9, 10 age group, all significant shifts were in a downward direction. The group with the lowest incidence of significant FSIQ score shift was the 13, 14 age group with only two subjects, one upward and one downward, outside the band of error at both the .95 and .99 levels of confidence.
The 13, 14 age group showed the broadest range of score changes, from -21 to +12. However, many of the score shifts were clustered in the lower numbers and so only a moderate rate of significant score shift was noted. The 15, 16 age group had the narrowest range of score shifts, from -11 to +6. The shifts were more evenly distributed across the range and a moderate rate of significant shift was seen.

Insert Table 10 about here

Regression to the Mean

Regression to the mean is a phenomenon that may account for some of the variance in changes in scores over time. This tendency for extreme scores to move toward the mean was examined in this study by comparing the means of the first and second administrations of the groups of subjects at the upper and lower extremes of the ranges of scores. Extreme scores were operationally defined as those scores outside +/-1 SD of the mean of the present group for each component IQ score and for the FSIQ. The means of the scores falling outside +/-1 S.D. of the mean and their differences are presented in Table 11. A large decline was seen at the upper extreme of the VIQ scores but very little upward movement was seen at the lower extreme. Since regression to the mean should be uniform between the upper and lower extreme scores when it is operating, this disparity between
the two directions of change indicates that regression to the mean was probably not a factor in the VIQ score shifts.

The same disparity was apparent in the FSIQ score shifts, so regression to the mean was probably not a factor in the changes in FSIQ scores.

The shifts in the means of the groups at the upper and lower extremes on the range of PIQ scores were relatively uniform compared to those of the VIQ and FSIQ. Therefore, regression to the mean may have accounted for some of the variance in the PIQ score shifts.

Discussion

Significant drops in scores across time were found by the repeated measures ANOVA for all three scales, VIQ, PIQ and FSIQ. Significant age differences were found in VIQ and PIQ. Significant interaction effects were found for VIQ and FSIQ, which means that significant differences in score shifts exist between age groups for both VIQ and FSIQ scores.

An examination of age group means revealed that age group 9, 10 exhibited the highest VIQ1 mean and that the same age group showed the largest change in group mean between the first and second administrations.
Individual changes observed in VIQ scores were found to be in concurrence with the data provided by the ANOVA procedures. An examination by group revealed that age group 9, 10 showed a considerably higher rate of significant VIQ score shifts than any of the other age groups. In addition, all significant shifts were in a negative direction.

A possible explanation of the significantly larger decrease in VIQ associated with the 9, 10 age group may be program related. The first administration of the VIQ would have occurred at ages 6 and 7 for this group when most of the subjects were beginning first and second grade. The age of this group at first administration is also a common time for initial referrals for psychoeducational evaluations and for special education resource programs to begin. This often means pullout from the regular classroom for portions of a day. It is possible that material missed by being out of the classroom for special programs may result in reduced development of skills sampled by the WISC-R.

Another possible explanation, advanced by Schmidt et al (1989) in citing Snider and Tarver (1987) is that around the third grade, students have typically mastered sufficient reading skills so that they are beginning to learn by reading, i.e. acquiring more information and vocabulary through reading experiences. Students with reading disabilities will not be able to access the information that other students typically access through reading and, unless
instruction is provided through some alternate method, the student may suffer an ever increasing deficit in acquired knowledge. At ages 9 and 10, when students are typically entering the fourth and fifth grades, the cumulative secondary effects of a reading disability may first begin to appear in reduced scores on WISC-R subtests that tap skills that are dependent on reading. While this explanation is a plausible one, it is unproven and to verify it will require additional research, including examination of age groups. Studies of special populations have not typically examined the effects of age groups.

The Pearson r procedure performed in the present study found adequate coefficients of stability for the VIQ and FSIQ scores between administrations. The Pearson r, however, does not examine changes in score as did the ANOVA techniques, but examines changes in the relative positions of the subjects from the first to the second administration. This means that the subjects maintained their relative position, to an acceptable degree, from the first to the second administration for VIQ and FSIQ, but that relative postions of subjects changed to a greater extent for PIQ. Test-retest correlation, however, is not an adequate measure of stability. Only relative position is examined by this method and if all the scores changed but the relative positions remained the same, the changes would go undetected and a high correlation coefficient would be observed.
Significant changes in scores could be masked if this procedure alone were used to examine stability.

An examination of the frequency distribution of changes in individual scores for the entire sample of 81, revealed high numbers of score shifts outside the band of error for VIQ, PIQ and FSIQ.

In contrast to the high number of score shifts in VIQ for the 9, 10 age group which were discussed above, age group 11, 12, showed no score shifts in VIQ at the 99% confidence level. At the 95% level, the rate of VIQ score shifts was only 11% of N, all of which were increases. The rate of VIQ score shifts at the 95% and 99% levels of confidence for age groups 13, 14 and 15, 16 were commensurate with the VIQ shifts observed in the entire sample.

The appearance of less diverse numbers of substantial PIQ score shifts between age groups is consistent with the findings of the repeated measures ANOVA. As discussed earlier, that procedure found no interaction effects between times of administration and age groups and, therefore, is in concurrence with the finding of no substantial differences in PIQ score changes between any of the age groups.

A greater number of decreased scores than increased scores was noted for the FSIQ distribution of score shifts
for the entire sample. This pattern is probably reflective of the extreme diversity seen in the VIQ pattern.

A comparison of the frequencies of significant changes in individual VIQ, PIQ and FSIQ scores between the present study and the reviewed studies which reported frequency data, revealed comparable rates of score shifts in general. An exception is the study by Schmidt et al (1989), which reported VIQ score shifts at a rate approximately twice that of the other reviewed studies, as well as that of the present study.

Regression to the Mean

The phenomenon of extreme scores regressing toward the mean on readministrations of the WISC-R was examined in order to determine whether or not it represented a significant source of variance in the present study. Telzrow (1990), citing Telzrow (1985), submits that this phenomenon may lead to overidentification of high-ability students while low-ability students may be underidentified. The results of the present study were probably not affected by regression to the mean.

Limitations

While the present study has an advantage over other studies in its examination by age groups, it is limited in that selection of a random sample was not possible. Randomization was not possible for a number of reasons including the limitations imposed by the selection criteria,
incomplete data, as well as other reasons which were discussed above. Since 80% of the LD population represented in the files was rejected, only a narrow sample of the LD population remained available for inclusion in the study. This will limit the generalizability of the findings.

Of those student files excluded from the study, approximately 25% were unusable because the scores of only one evaluation were present. We might infer that those students transferred to other school districts and, as suggested by Stavrou (1990), may represent a group that is more transient than the group that was studied. It follows that the characteristics of that group may preclude generalizing current findings to transient students.

An additional limitation lies in the fact that the current sample was drawn from a rural population in a small geographic area. The results may not generalize to more densely populated areas, metropolitan areas or other geographic locations.

Finally, a limitation of the significant findings attributed to the 9, 10 age group must be viewed tentatively because of the small size of this group, consisting of only 12 subjects. Further research in this specific area would necessitate studying larger samples.

Suggestions for Research

The foregoing discussion of the hypothesis of secondary effects of reading disabilities as advanced by Schmidt et al
(1988) presents a plausible explanation for the decline in VIQ scores seen between ages 6, 7 and 9, 10, when the cumulative effects of a lack of information input via reading may first appear. Additional research is needed to further test this hypothesis. A possible research question might ask if declines in VIQ are seen in students with reading disabilities between ages 6, 7 and 9, 10 and, if so, are the declines significantly greater than those seen in other age groups.

To further test the hypothesis of secondary effects of reading disabilities, the same question might be applied to different cognitive levels within the reading disabled population. Differences in rate of change in WISC-R scores were noted in studies reviewed herein between mentally handicapped and LD subjects. It is possible that similar differences may appear between different cognitive levels within the LD population. It may be that students with higher levels of ability are better able to use context cues and other compensatory techniques to expand their vocabulary skills and knowledge of the world, thereby minimizing declines in VIQ between administrations.

**Educational Implications**

The present study has found sufficient test-retest differences between WISC-R scores to conclude that we cannot assume that scores will remain stable between three year evaluations of LD students. Shifts in IQ scores may reveal
important information about a student's progress and about the effectiveness of educational programs. Therefore, it follows that retesting with the WISC-R and other measures including achievement tests is necessary in reevaluations.

Ross-Reynolds (1990) proposes a model for reevaluations which focuses on assessing student progress and program effectiveness. Standardized tests may be used for summative assessment of overall gains, while curriculum based measurement (CBM) is recommended for more accurate measurement of short term gains. Other data gathering methods are also employed including a review of the Individualized Education Program (IEP); parent, teacher and student interviews; and classroom observations. CBM could serve as a problem solving technique in the classroom and would provide a consulting role for the school psychologist.

The instability of WISC-R scores of LD students that has been demonstrated by the present study as well as a number of other studies suggests that use of a discrepancy formula may also have insufficient reliability in initial LD evaluations. It is suggested that the initial verification of a learning disability incorporate a broader range of data than sole reliance on a discrepancy between IQ and achievement scores. Sole use of a discrepancy formula in identifying a learning disability is prohibited by federal law (Ross-Reynolds, 1990). However, many states, including Nebraska, utilize it in the decision making process as a
WISC-R Score Stability

necessary condition. A more comprehensive approach to the verification of learning disabilities has been suggested by Ferguson and Mamen (1985). This approach emphasizes complete history taking including family, school, medical and developmental; parent information including their perceptions and expectations of the child; testing, including intelligence, achievement, auditory-language and visual-motor-spatial; and behavioral observations. The authors support all of this data entering into the verification of the learning disability, rather than relying only on a discrepancy formula, which gives little specific information about the type of learning disability or the program needs.

The findings of significant differences in the scores of the 9, 10 age group discussed above suggest implications for educational programming. The significant decline in VIQ scores observed in this group may be a reflection of the secondary effects of reading disabilities. As reading becomes increasingly important as a medium for acquiring information and vocabulary, effective instruction for reading disabled students becomes especially important. School psychologists need to be alert to indicators of secondary effects of reading disabilities and make appropriate recommendations for programming. Recommendations might include direct instruction, a behaviorally based approach which includes direct or
criterion-referenced measurement of the skills required for reading and language performance. Utilization of this system in a study by Lloyd et al (1981) was shown to be effective in making significantly greater gains than other methods in improving LD students' reading and language skills.

In conclusion, the present study has not demonstrated the WISC-R to have adequate test-retest stability over a three year interval in a population of LD students. It follows that retesting for triennial evaluations is necessary both from the standpoint of confirming the existence of a discrepancy and of gaining important information relating to student progress and educational programming. Intelligence testing using the WISC-R should be an important part of alternative approaches to LD evaluations.
Table 1

Highlights of Studies Examining WISC-R Score Stability between Reevaluations of Learning Disabled Students

<table>
<thead>
<tr>
<th>Study</th>
<th>Adequate stability</th>
<th>Statistics</th>
<th>Retest necessary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anderson et al (1989)</td>
<td>No</td>
<td>Pearson r, t-test</td>
<td>Yes</td>
</tr>
<tr>
<td>Oakman &amp; Wilson (1988)</td>
<td>Yes</td>
<td>Correlated t-test, ANOVA</td>
<td>No</td>
</tr>
<tr>
<td>Schmidt et al (1989)</td>
<td>Yes</td>
<td>Correlated t-test, ANOVA</td>
<td>NA^b</td>
</tr>
<tr>
<td>Stavrou (1990)</td>
<td>No</td>
<td>Frequency distribution</td>
<td>Yes</td>
</tr>
<tr>
<td>Webster (1988)</td>
<td>No</td>
<td>Frequency distribution</td>
<td>Yes</td>
</tr>
</tbody>
</table>

^aOpinion of the author

^bNot addressed
### Table 2

Comparison of Percentages of Significant Shifts in IQ Scores in Studies Presenting Frequency Data

<table>
<thead>
<tr>
<th>Study</th>
<th>VIQ</th>
<th>PIQ</th>
<th>FSIQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schmidt et al(^a)</td>
<td>53%</td>
<td>22%</td>
<td>53%</td>
</tr>
<tr>
<td>Stavrou(^b) (Time 1 to 2)</td>
<td>27%</td>
<td>24%</td>
<td>33%</td>
</tr>
<tr>
<td>(Time 2 to 3)</td>
<td>23%</td>
<td>18%</td>
<td>23%</td>
</tr>
<tr>
<td>Webster(^c)</td>
<td>23%</td>
<td>23%</td>
<td>29%</td>
</tr>
</tbody>
</table>

\(^{a}\) \(_{p}<.05\), change in VIQ= +/-7, change in PIQ= +/-9, change in FSIQ= +/-6

\(^{b}\) \(_{p}<.05\), change in VIQ= +/-10, change in PIQ= +/-13, change in FSIQ= +/-9

\(^{c}\) \(_{p}<.05\), change in VIQ= +/-7, change in PIQ= +/-10, change in FSIQ= +/-6
Table 3

Means and Standard Deviations of First and Second Administrations of WISC-R VIQ, PIQ and FSIQ

<table>
<thead>
<tr>
<th>Age group</th>
<th>N</th>
<th>VIQ1 M</th>
<th>VIQ1 S.D.</th>
<th>VIQ2 M</th>
<th>VIQ2 S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>9, 10</td>
<td>12</td>
<td>99.000</td>
<td>9.035</td>
<td>90.417</td>
<td>10.388</td>
</tr>
<tr>
<td>11, 12</td>
<td>18</td>
<td>95.278</td>
<td>11.676</td>
<td>96.722</td>
<td>10.431</td>
</tr>
<tr>
<td>13, 14</td>
<td>24</td>
<td>90.292</td>
<td>12.231</td>
<td>89.750</td>
<td>11.117</td>
</tr>
<tr>
<td>15, 16</td>
<td>27</td>
<td>87.444</td>
<td>9.124</td>
<td>84.926</td>
<td>10.099</td>
</tr>
<tr>
<td>Total</td>
<td>81</td>
<td>91.741</td>
<td>11.307</td>
<td>89.790</td>
<td>11.202</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age group</th>
<th>N</th>
<th>PIQ1 M</th>
<th>PIQ1 S.D.</th>
<th>PIQ2 M</th>
<th>PIQ2 S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>9, 10</td>
<td>12</td>
<td>103.583</td>
<td>18.048</td>
<td>98.083</td>
<td>17.207</td>
</tr>
<tr>
<td>11, 12</td>
<td>18</td>
<td>108.444</td>
<td>7.801</td>
<td>107.000</td>
<td>11.256</td>
</tr>
<tr>
<td>13, 14</td>
<td>24</td>
<td>100.375</td>
<td>11.485</td>
<td>99.417</td>
<td>11.504</td>
</tr>
<tr>
<td>Total</td>
<td>81</td>
<td>105.037</td>
<td>11.584</td>
<td>102.951</td>
<td>12.177</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age group</th>
<th>N</th>
<th>FSIQ1 M</th>
<th>FSIQ1 S.D.</th>
<th>FSIQ2 M</th>
<th>FSIQ2 S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>9, 10</td>
<td>12</td>
<td>101.083</td>
<td>12.347</td>
<td>93.333</td>
<td>12.687</td>
</tr>
<tr>
<td>11, 12</td>
<td>18</td>
<td>101.056</td>
<td>9.052</td>
<td>101.222</td>
<td>9.723</td>
</tr>
<tr>
<td>13, 14</td>
<td>24</td>
<td>94.583</td>
<td>10.274</td>
<td>93.583</td>
<td>10.375</td>
</tr>
<tr>
<td>15, 16</td>
<td>27</td>
<td>96.296</td>
<td>8.475</td>
<td>93.815</td>
<td>9.060</td>
</tr>
<tr>
<td>Total</td>
<td>81</td>
<td>97.556</td>
<td>9.994</td>
<td>95.321</td>
<td>10.485</td>
</tr>
</tbody>
</table>
Table 4

Repeated Measures Analysis of Variance Table for First and Second Administrations of VIQ by Age Groups

<table>
<thead>
<tr>
<th>SV</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Subjects</td>
<td>80</td>
<td>18,463.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age Groups (A)</td>
<td>3</td>
<td>2,519.00</td>
<td>839.07</td>
<td>4.06</td>
<td>.01</td>
</tr>
<tr>
<td>S/A</td>
<td>77</td>
<td>15,944.09</td>
<td>.207.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Subjects</td>
<td>81</td>
<td>2,006.93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIQ Time (B)</td>
<td>1</td>
<td>239.03</td>
<td>239.03</td>
<td>13.07</td>
<td>.001</td>
</tr>
<tr>
<td>AB</td>
<td>3</td>
<td>359.87</td>
<td>131.96</td>
<td>7.22</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>BS/A</td>
<td>77</td>
<td>1,408.03</td>
<td>18.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>161</td>
<td>20,470.02</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 5

Repeatead Measures Analysis of Variance Table for First and Second Administrations of PIQ by Age Groups

<table>
<thead>
<tr>
<th>SV</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Subjects</td>
<td></td>
<td>19,857.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age Groups (A)</td>
<td>3</td>
<td>1,900.63</td>
<td>633.54</td>
<td>2.72</td>
<td>.050</td>
</tr>
<tr>
<td>S/A</td>
<td>77</td>
<td>17,956.87</td>
<td>233.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Subjects</td>
<td></td>
<td>2,964.54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIQ Time (B)</td>
<td>1</td>
<td>225.34</td>
<td>225.34</td>
<td>6.55</td>
<td>.012</td>
</tr>
<tr>
<td>AB</td>
<td>3</td>
<td>89.00</td>
<td>29.67</td>
<td>.86</td>
<td>.465</td>
</tr>
<tr>
<td>BS/A</td>
<td>77</td>
<td>2,650.20</td>
<td>34.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>161</td>
<td>22,822.04</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 6

**Repeated Measures ANOVA Table for First and Second Administrations of FSIQ by Age Group**

<table>
<thead>
<tr>
<th>SV</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Subjects</td>
<td>80</td>
<td>15,260.38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age Groups (A)</td>
<td>3</td>
<td>1,179.12</td>
<td>393.04</td>
<td>2.15</td>
<td>.101</td>
</tr>
<tr>
<td>S/A</td>
<td>77</td>
<td>14,081.26</td>
<td>182.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Subjects</td>
<td>81</td>
<td>1,806.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSIQ Time (B)</td>
<td>1</td>
<td>281.33</td>
<td>281.33</td>
<td>17.03</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>AB</td>
<td>3</td>
<td>253.53</td>
<td>84.51</td>
<td>5.12</td>
<td>.003</td>
</tr>
<tr>
<td>BS/A</td>
<td>77</td>
<td>1,271.75</td>
<td>16.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>161</td>
<td>17,066.99</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 7

**Frequency of Significant Shifts in VIQ, PIQ and FSIQ Scores Between Administrations Based on Bands of Error**

<table>
<thead>
<tr>
<th></th>
<th>Number of + and - Shifts</th>
<th>Percentage of N</th>
<th>Number of + Shifts</th>
<th>Number of - Shifts</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25 *</td>
<td>31 *</td>
<td>6 *</td>
<td>19 *</td>
</tr>
<tr>
<td></td>
<td>15 **</td>
<td>19 **</td>
<td>2 **</td>
<td>13 **</td>
</tr>
<tr>
<td>PIQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18 *</td>
<td>22 *</td>
<td>6 *</td>
<td>12 *</td>
</tr>
<tr>
<td></td>
<td>10 **</td>
<td>12 **</td>
<td>5 **</td>
<td>5 **</td>
</tr>
<tr>
<td>FSIQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>21 *</td>
<td>26 *</td>
<td>4 *</td>
<td>17 *</td>
</tr>
<tr>
<td></td>
<td>15 **</td>
<td>19 **</td>
<td>2 **</td>
<td>13 **</td>
</tr>
</tbody>
</table>

*aKaufman, 1979

* 95% level of confidence

** 99% level of confidence
Table 8

**Frequency of Significant Shifts in VIQ Scores by Age Group**

*Based on Bands of Error*[^1]

<table>
<thead>
<tr>
<th>Age Group</th>
<th>N</th>
<th>Number of + and - Shifts</th>
<th>Percentage of N</th>
<th>Number of + shifts</th>
<th>Number of - shifts</th>
</tr>
</thead>
<tbody>
<tr>
<td>9,10</td>
<td>12</td>
<td>8 *</td>
<td>67 *</td>
<td>0 *</td>
<td>8 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8 **</td>
<td>67 **</td>
<td>0 **</td>
<td>8 **</td>
</tr>
<tr>
<td>11,12</td>
<td>18</td>
<td>2 *</td>
<td>11 *</td>
<td>2 *</td>
<td>0 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 **</td>
<td>0 **</td>
<td>0 **</td>
<td>0 **</td>
</tr>
<tr>
<td>13,14</td>
<td>24</td>
<td>8 *</td>
<td>33 *</td>
<td>3 *</td>
<td>5 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 **</td>
<td>17 **</td>
<td>1 **</td>
<td>3 **</td>
</tr>
<tr>
<td>15,16</td>
<td>27</td>
<td>7 *</td>
<td>26 *</td>
<td>1 *</td>
<td>6 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 **</td>
<td>11 **</td>
<td>1 **</td>
<td>2 **</td>
</tr>
</tbody>
</table>

[^1]: Kaufman, 1979

* 95% level of confidence

** 99% level of confidence
WISC-R Score Stability

Table 9

Frequency of Significant Shifts in PIQ Scores by Age Group
Based on Bands of Error<sup>a</sup>

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Number of N + and - Shifts</th>
<th>Percentage of N</th>
<th>Number of + shifts</th>
<th>Number of - shifts</th>
</tr>
</thead>
<tbody>
<tr>
<td>9,10</td>
<td>12</td>
<td>3 *</td>
<td>25 *</td>
<td>1 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 **</td>
<td>17 **</td>
<td>1 **</td>
</tr>
<tr>
<td>11,12</td>
<td>18</td>
<td>4 *</td>
<td>22 *</td>
<td>2 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 **</td>
<td>11 **</td>
<td>1 **</td>
</tr>
<tr>
<td>13,14</td>
<td>24</td>
<td>9 *</td>
<td>38 *</td>
<td>4 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 **</td>
<td>25 **</td>
<td>4 **</td>
</tr>
<tr>
<td>15,16</td>
<td>27</td>
<td>3 *</td>
<td>11 *</td>
<td>0 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 **</td>
<td>4 **</td>
<td>0 **</td>
</tr>
</tbody>
</table>

<sup>a</sup>Kaufman, 1979

* 95% level of confidence

** 99% level of confidence
Table 10

Frequency of Significant Shifts in FSIQ Scores by Age Group Based on Bands of Error\(^a\)

<table>
<thead>
<tr>
<th>Age Group</th>
<th>N</th>
<th>+ and - Shifts</th>
<th>Percentage of N</th>
<th>Number of + shifts</th>
<th>Number of - shifts</th>
</tr>
</thead>
<tbody>
<tr>
<td>9,10</td>
<td>12</td>
<td>8 *</td>
<td>67 *</td>
<td>0 *</td>
<td>8 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 **</td>
<td>58 **</td>
<td>0 **</td>
<td>7 **</td>
</tr>
<tr>
<td>11,12</td>
<td>18</td>
<td>2 *</td>
<td>11 *</td>
<td>1 *</td>
<td>1 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 **</td>
<td>11 **</td>
<td>1 **</td>
<td>1 **</td>
</tr>
<tr>
<td>13,14</td>
<td>24</td>
<td>8 *</td>
<td>33 *</td>
<td>4 *</td>
<td>4 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 **</td>
<td>17 **</td>
<td>1 **</td>
<td>3 **</td>
</tr>
<tr>
<td>15,16</td>
<td>27</td>
<td>7 *</td>
<td>26 *</td>
<td>1 *</td>
<td>6 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 **</td>
<td>11 **</td>
<td>0 **</td>
<td>3 **</td>
</tr>
</tbody>
</table>

\(^a\)Kaufman, 1979

* 95% level of confidence
** 99% level of confidence
Table 11

Comparison of Mean Scores from Time 1 to Time 2 of Individuals at the Upper and Lower Extremes

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Group Mean</th>
<th>Time 1</th>
<th>Group Mean</th>
<th>Time 2</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIQ</td>
<td>&lt;-1 S.D.</td>
<td>9</td>
<td>71.667</td>
<td>72.556</td>
<td>+.889</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;+1 S.D.</td>
<td>14</td>
<td>115.427</td>
<td>107.071</td>
<td>-8.358</td>
<td></td>
</tr>
<tr>
<td>PIQ</td>
<td>&lt;-1 S.D.</td>
<td>9</td>
<td>82.556</td>
<td>86.444</td>
<td>+3.888</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;+1 S.D.</td>
<td>8</td>
<td>123.750</td>
<td>118.875</td>
<td>-4.875</td>
<td></td>
</tr>
<tr>
<td>FSIQ</td>
<td>&lt;-1 S.D.</td>
<td>12</td>
<td>82.417</td>
<td>81.000</td>
<td>-1.417</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;+1 S.D.</td>
<td>16</td>
<td>111.938</td>
<td>106.375</td>
<td>-5.563</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\text{At first administration}\)
References


