THE EFFECTIVENESS OF CULTURE-FREE TESTS IN MEASURING THE INTELLECTUAL CHARACTERISTICS OF GERMAN IMMIGRANTS TO CANADA

by

EDWARD NORMAN ELLIS

A THESIS

submitted to

OREGON STATE COLLEGE

in partial fulfillment of
the requirements for the
degree of

DOCTOR OF EDUCATION

June 1956

APPROVED:

Redacted for Privacy

Associate Professor of Psychology.

In Charge of Major

Redacted for Privacy

Dean of School of Education

Redacted for Privacy

Chairman of School Graduate Committee

Redacted for Privacy

Dean of Graduate School

Date thesis is presented April 27, 1956.

Typed by Winifred May Ellis.

ACKNOWLEDGMENT

The writer is indebted to the members of his committee, Dean F. R. Zeran, Dean H. P. Hansen, Dr. J. W. Sherburne, Dr. R. R. Reichart, Professor S. E. Williamson, Dr. F. W. Decker, and Dr. W. R. Crooks, for their kindness and helpful guidance through his graduate program. He is grateful to Dr. Reichart for his inspiration and assistance in the planning of the program. Dr. Crooks, as the professor in charge of the major, has supervised the preparation of the thesis. Particular thanks are expressed to him for constant encouragement, meaningful criticisms and helpful suggestions. His keen perception of problems made difficult tasks easy, and his approach made them pleasant. The writer is deeply indebted to Dr. J. E. R. Li, Professor of Mathematics, for his invaluable assistance with the statistical aspects of the study.

In the course of selecting instruments for this research, communications were exchanged with the following persons: Dr. A. Schwarzlose, Berlin Lichterfelde, Dr. H. M. Ledig, Schulpsychologische Arbeitstelle, Berlin, Professor John C. Raven, The Crichton Royal, Dumfries, Scotland, Professor R. W. Payne, Maudsley Hospital, London, Dr. Grace Arthur, St. Paul, Minnesota, Dr. James H. Ricks, Psychological Corporation, New York, and Dr. Raymond B.

Cattell, University of Illinois, Urbana, Illinois. The information that they provided was most useful and it is gratefully acknowledged.

The writer is grateful to Dr. R. F. Sharp, superintendent of schools in Vancouver, for permission to conduct the investigation, to the principals and teachers for their excellent cooperation, and to the girls and boys who, as model subjects, cheerfully submitted to the tests.

Finally, the writer expresses his appreciation of the patient and faithful help of his wife, Winifred. Without her unfailing encouragement and the contribution of her typing talent, there would not have been a thesis.

TABLE OF CONTENTS

THE EFFECTIVENESS OF CULTURE-FREE TESTS IN MEASURING THE INTELLECTUAL CHARACTERISTICS OF GERMAN IMMIGRANTS TO CANADA

CHAPTER I

PRESENTATION OF THE PROBLEM

Introduction

Canada has long been a melting pot of many ethnic groups. During 1954 alone, an average of more than four hundred immigrants, representing nearly fifty different nationalities, entered the Dominion daily (34,p.1). Large numbers of these settlers have moved to western Canada where the children enter the public schools.

The satisfactory assimilation of immigrant children particularly depends upon the kind of schooling they receive in their new homeland. The education of these new Canadians can be successful only to the extent that provision is made for their individual differences. Thus, there is a need for a valid measure of their intellectual characteristics so that instruction may be appropriate to the level of their native talents.

Many of the mental tests currently used in Canadian schools are unsuitable instruments for immigrant children. Most of them are dependent upon the subject's

facility with the English language and his familiarity with American culture. Consequently, these tests have little validity for anyone who is deficient in the use of the English language or foreign to the American way of life. To overcome these difficulties, test-constructors have designed nonverbal "culture-free" tests. Some of these are individual tests while others are administered in a group situation. The tests vary widely in form, length, material and procedure. Some are done with pencil and paper, others consist of manipulative tasks. Many "culture-free" tests are expressed in pictures or cartoons, while others use symbols, digits, designs, blocks, objects and form boards. It is probable that these tests are not all equally effective in removing cultural contamination from intelligence testing. It is with these instruments and their use among immigrant Canadians that this project is concerned.

The Problem

This experiment has been designed to determine the extent to which "culture-free" tests succeed in minimizing cultural factors likely to affect the measurement of the intelligence of immigrant German children entering Canadian schools. Comparisons will be made of selected culture-free instruments (and their sub-tests) in the hope

that some conclusions may be drawn about the kind of instrument that might be a valid, appropriate and convenient measure of the scholastic aptitudes of immigrant children to Canada.

At this point it should be made clear that this study is not intended in any way to be a survey of national differences in intellectual ability. The immigrants have not been drawn proportionately from all socioeconomic and educational levels and they cannot be assumed to be representative of their national group. The relative performance of Germans and Canadians is not, then, the concern of this study; it is the nature and effectiveness of the instruments being used.

The Need for this Investigation

During recent years, immigration into Canada has been accelerated. Some 154,227 persons were admitted during 1954, and of these 29,845 were of German origin, constituting the largest national group of immigrants to Canada (see table I, Appendix B). These "New Canadians" migrate to all of the provinces (see table II, Appendix B). During 1954, 12,197 persons settled in British Columbia and a large proportion of these have established their homes in greater Vancouver. The public schools of the city have been challenged to meet the needs of these

people. The extent to which the school facilities have been strained can be realized when it is remembered that of the German immigrants who entered Canada last year, 28 per cent were under the age of eighteen years (see table III, Appendix B).

Few of these immigrants are proficient in the use of oral or written English. To meet their needs, classes in "English for New Canadians" have been established. Adult immigrants go to these classes in the night schools while children attend the day schools. Those of primary age (6, 7, and 8 years) usually enroll in the primary grade and learn English while attending regular classes with other Canadian children of their age. The older children attend the special English classes for periods of from four to ten months. In Vancouver, the first of these special "English for New Canadians" classes for older children was organized in 1949 in the Lord Strathcona School. Within three years, twenty-six classes in various schools of the city had been established to serve the needs of nearly four hundred immigrant children. However, during the past year there has been a decrease and at the time of writing only thirteen classes (with an enrollment of 267) are in operation.

Considerable attention has been given to the preparation of the course of study for these classes.

Materials and text-books have been carefully prepared and graded. Special techniques are being employed and the teachers for these classes are very carefully selected. All who are concerned with the teaching of English to New Canadians realize that the program can be successful only to the extent that it provides for individual differences, and this can be done only when the instructor knows the capacities of his students.

Many of these immigrants are of such an age that they cannot afford to be in school for more than a very few years. It is highly important that their time in school be used economically. An effective testing program will help to indicate for each individual his strengths and weaknesses and will aid in pointing out the way in which his time and efforts may be put to best advantage.

Further, these New Canadians are seeking assistance in personal, vocational, and educational matters. To meet this need, counselors are compiling files of personal data for each of these individuals. A valid measure of the immigrant's mental characteristics is of prime value in such an inventory.

In order, then, that the instruction and counseling of these new Canadians may be more effective, there is a need for suitable mental testing. At the present time,

only a few intelligence tests are being administered. The validity of the results on these tests is questionable because of differences in language and cultural background. These factors are said to be minimized in many of the new "culture-free" tests. It seems important and timely to the writer that there should be an investigation into the relative worth and effectiveness of these tests with German immigrants to Canada to the end that we may learn more about the uses and limitations of these tests, and, if possible, discover how effective these tests are in giving scores that are valid and reliable.

Aims

The following are specific questions that, it is hoped, shall be answered by this experiment:

(1) Do immigrant children (who have been matched with second generation Canadians on the bases of age, sex and raw score on the Progressive Matrices Test) score consistently lower than, higher than, or about the same as Canadians on other tests used in this research?

(2) If a difference exists between the performance of the immigrant children and that of Canadians, for which test is this difference greatest? Or, are the nationality differences about the same for all of the tests?

(3) Is the average score for boys higher than, lower than or about the same as the average score for girls?

(4) Which of the two national groups has the greater variation due to sex?

(5) Is the variation in scores due to nationality greater for boys or girls?

(6) For boys, which tests gives a mean score not significantly different from those on other tests?

(7) For boys, which of the tests would give about the same relative ranking in a large unselected population as that given to the same group by the Progressive Matrices test?

(8) Which test, if any, would give a mean score for girls that is not significantly different from the mean score on other tests?

(9) For girls, which of the tests would give about the same relative ranking in a large unselected population as that given to the same group by the Progressive Matrices test?

(10) Which, if any, of the "culture-free" tests gives a mean score that is significantly different from the mean score of the culture test?

(11) For which test is the difference the greatest between the mean score for Canadians and the mean

score for Germans?

(12) If there are components that operate in culture tests to make the relative standings of the national groups different from their relative standings on the Progressive Matrices test, do these test components operate to a greater or to a lesser extent in the Cattell, Wechsler and Arthur tests?

(13) What are the Pearsonian coefficients of correlation between results on any two of the tests for boys and for girls of both national groups, and are these values significant?

(14) Between which two tests is there the greatest agreement?

(15) For Canadians, is the mean score on subtests of the Wechsler-Bellevue Intelligence Test higher than, lower than, or the same as the mean score for Germans?

(16) How does the mean score for boys on the subtests of the Wechsler-Bellevue Intelligence Test compare with the mean score for girls?

(17) On which subtest of the Wechsler-Bellevue Intelligence Test did the groups perform best?

(18) On which subtest of the Wechsler-Bellevue Intelligence Test did the groups tend to perform poorly?

(19) For which of these subtests was the superior-

ity in performance of Canadians over Germans most marked?

(20) For which of these subtests was the superiority in the performance of Canadians over Germans least marked?

(21) For which of these subtests was the superiority in the performance of boys over girls most marked?

(22) On which subtest did the girls excell the boys?

(23) For Canadians, is the average score on the Arthur Point Scale of Performance Tests (Revised Form II) higher than, lower than, or the same as the average score for Germans?

(24) How does the average score of boys on the Arthur Point Scale of Performance Tests (Revised Form II) compare with the average score for girls?

(25) On which subtest of the Arthur Point Scale did the groups perform best?

(26) On which subtest of the Arthur Point Scale did the groups tend to perform poorly?

(27) For which of these subtests was the superiority in performance of Canadians over Germans most marked?

(28) For which of these subtests was the superiority in the performance of Canadians over Germans

least marked?

(29) For which of these subtests was the superiority in the performance of boys over girls most marked?

(30) On which subtests did the girls excell the boys?

(31) Which subtest displayed the greatest discriminative power?

(32) On which subtest of the Cattell test did the groups perform best?

(33) On which subtest of the Cattell test did the groups perform poorly?

(34) On which subtest of the Cattell test were nationality differences greatest?

(35) On which subtest of the Cattell test were sex differences greatest?

(36) How well do the test results agree with teachers' ratings?

(37) How do the tests compare relatively in their ability to predict teacher-assigned grades for immigrant children?

(38) How do the tests compare relatively in their ability to predict teacher-assigned ratings for Canadian students?

Definitions

Before progressing further, it seems desirable to establish a common understanding of certain basic expressions that will be employed in this study. There appears to be a need to define the following terms:

Bilinguist - A person brought up in a family where two languages are used interchangeably.

Canadian - For the purposes of this study, a Canadian is a person born and raised in Canada, whose parents were born and raised in Canada, and in whose home English is the primary language.

Culture - The integrated customs, language, traditions, beliefs, habits, morals, and social forms of a group of persons; the man-made environment of one society that distinguishes it from another social group.

"Culture-Fair" - A test is "culture-fair" if it measures fairly the basic problem-solving ability of children from different cultural backgrounds. While cultural materials are not eliminated, the design of the test is such as to remove any bias that would operate in favor of the children of one particular culture. Generally speaking, the items deal with materials equally common to the various groups on which the test is to be used. The language and symbols employed in the test are equally familiar to all subjects and the test is so organized and administered as to stimulate equal degrees of interest and motivation for children from different cultures.

"Culture-Free" - Literally, "culture-free" implies that the test measures intelligence entirely independent of cultural experiences. To construct items for such a test so that the subject will make responses that will be free from cultural influences is obviously a difficult, if not an impossible task. Even the nature of a child's response to culturally novel materials is likely to be conditioned by patterns of habits and attitudes that are themselves culturally-

determined.

In practice, the term "culture-free" is commonly used to refer to those mental tests that attempt to minimize cultural factors and that, in so far as it is possible, permit them to operate to the same extent for each of the cultural groups being tested. Here, the term "culture-free" is synonymous with "culture-fair" and it is in this sense that it will be employed in this study.

Culture-test - A test consisting of items that are drawn from a particular cultural environment and that will be biased in favor of one cultural group over another.

'g' factor - The general intelligence factor which, according to Spearman's two-factor theory, is fundamental to all correlated abilities for the same individual as distinguished from the specific factors which vary in different activities.

German - Those New Canadians who emigrated from Germany and who are considered herein as "New Canadians".

Group test - A test administered to several subjects simultaneously by a single examiner.

Immigrant - Will be used synonymously with "New Canadian".

Individual test - A test administered to one person at a time.

Intelligence - The capacity of an individual to meet new situations quickly and successfully; the capacity to meet a novel situation by improvising a novel adaptive response; problem-solving ability.

Monoglot - A person brought up in a family where only one language is used.

"New Canadian" - An immigrant who has recently arrived in Canada from a foreign country where English is not the primary language, who has not been

in the Dominion longer than six months at the time of testing, and whose understanding of the English language is such that he has been enrolled in a special English language class.

Nonlanguage test - A type of test in which the instructions are given by pantomime and no words are required in solving the test.

Nonverbal test - A type of mental test in which no words are used in the test content, but the instructions may be given verbally or by pantomime.

Performance scale - A series of performance tests in which the exercises are arranged in order of increasing difficulty.

Performance test - A type of mental test in which the role of language is greatly diminished, the test material consisting of concrete objects instead of words, and the responses consisting of manipulations of these objects though the directions are often given verbally.

Limitations

There are always restrictions that have to be imposed when conducting research. The following appear to be the most significant limitations of this experimental study and they should be carefully considered in attempting to understand, to evaluate, and to interpret the findings of this investigation.

(1) The influence of many variable factors on psychological test performance.

It should be pointed out that individuals who differ in racial affiliation also differ in many other respects. Klineberg (100, p.152) lists the following as

factors that affect "racial differences", language, schooling, culture, socioeconomic status, rapport, motivation, sampling and speed. To a large extent, language differences limit the validity of comparisons made between racial groups. Differences also exist in the national economic levels, in the general social conditions, in the opportunities for education, in the facilities for training, and in the traditional and cultural backgrounds of the groups. Under these variable conditions the emotional attitudes, interests, ideals, and preferences of the two groups are not likely to be the same.

(2) Difficulties of communication.

There are obvious difficulties that arise when an examiner of one nationality administers tests to subjects of another national group. Although the extent of this handicap in the present study is not definitely known, the writer has endeavored to maintain good rapport, to provide a permissive atmosphere that would minimize any anxiety of the subject, and to be reasonably certain that his instructions were at all times understood.

(3) Failure to measure the same traits.

Whenever a language difficulty invalidates the use of verbal tests, the range of processes that can be measured in that group is greatly restricted. In non-language and performance tests, according to Anastasi

and Foley (7,pp.486-487), spatial aptitude plays the dominant role while most paper-and-pencil tests measure chiefly verbal ability and to a slighter extent, numerical ability. In the present study, it is probable that the tests are not tapping the same abilities nor are any two tests likely to depend on the same special ability to exactly the same extent.

(4) The questionable suitability of norms.

One fundamental problem in the interpretation of the test results centers in the choice of standards for the evaluation of diverse peoples. The findings of this study are limited to the extent that the norms used are appropriate for both groups.

(5) The size of the sample.

For practical reasons, the size of each of the groups being tested was restricted to fifty persons.

(6) The selection of the members of the experimental group.

No attempt was made to randomize the sample or to obtain a group that would be representative (on a socio-economic, educational or any other basis) of the German national and cultural group. Immigrant children were enrolled in the experimental group in the order in which they were admitted to the special English class for New Canadians. It happened, quite by chance, that the boys

included in the experimental group exceeded the girls in mean score on the Progressive Matrices, which was the test used for matching purposes. This is an unfortunate circumstance in that the nature of the sample contributes to sex-variation in test results.

(7) Elapsed time between arrival and testing.

Although the writer made it his prime concern to administer the tests as promptly as possible after the immigrant's arrival so that experience with Canadian culture would be minimized, it was not always possible to do so. Some individuals arrived during school vacation period, others were slow to make registration for school, and for all of the immigrants there was a long transcontinental journey from the Atlantic sea board. While this period for most of the subjects was only a matter of a few weeks, the writer submits that it provides a source for error.

(8) Inequalities in motivation.

In administering any test, the examiner seeks the optimum performance of his subjects and seldom can he be certain that they are all equally well motivated. While efforts were made to put the subjects at ease, to establish rapport, to promote interest, and to stimulate maximum effort, it would be presumptuous to consider that these endeavors were completely successful.

(9) The restricted selection of instruments.

An attempt was made to use tests of different types and in some cases the selection was made from a relatively small number. Notably, the Begabungstest B-1 is one of very few German intelligence tests that have been widely used and well-standardized. This restricted selection may be one of the more serious limitations of this study.

(10) The restricted age-range of the subjects.

Because of the fact that most mental tests are designed for a particular age-range the subjects selected for this study were restricted to those of ten to fifteen years of age.

(11) The basis of matching.

The writer had to choose between two courses in designing the experiment, -- (a) to match the groups on the basis of scores on two different but well standardized verbal culture-tests (one in German, the other in English) and then to experiment only with various "culture-free" instruments; or (b) to match the groups on the basis of raw scores of one "culture-free" test that is currently being used with both national groups, and then to include the scores on culture tests with the other experimental data. While there is much to commend the first approach, the writer felt that the second would provide a sounder

basis for matching and this was the plan that was followed. One of the most serious limitations is that in this design the Progressive Matrices test is removed from the comparisons being made.

(12) The preponderance of 'C' grades assigned by teachers.

In the attempt, herein, to see which test gives results that agree most closely with teacher-assigned grades, the statistical result is unsatisfactory because a majority of the students were give 'C' standing and, relatively, few were placed in the extreme 'A' or 'E' categories.

(13) Restrictions of time and place.

The subjects that constituted the experimental group were enrolled in 'New Canadian' classes in Vancouver during the fall and winter of 1954. The control group consisted of second-generation Canadians attending regular classes in Vancouver during the same school term. These confinements of the groups with respect to time and place were made for the sake of convenience and expediency.

CHAPTER II

REVIEW OF EARLIER INVESTIGATIONS AND RELATED RESEARCH

Research that is related to the present study falls naturally into two divisions; (1) the refinement of psychometric instruments and the development of culture-free tests, and (2) racial comparisons of mental ability.

The Refinement of Psychometric Instruments and the Development of Culture-free tests.

There have been numerous investigations into the validity and effectiveness of intelligence tests. These studies have been designed with the aim of improving the tests so that they will do a better job of predicting future success. Generally speaking, such research includes many contemporary studies and bears directly on the present investigation.

While most of the research on culture-free tests has been conducted during the last ten years, the development of tests that would minimize cultural factors has extended over a much longer period. The history of

nonlanguage performance tests probably begins with Itard's use of a form board in 1801 to train his "wild-boy" and, subsequently, to identify mental defectives. Many simple sensori-motor tests were developed and employed during the latter part of the nineteenth century largely due to the influence of Wundt's laboratory of experimental psychology at Leipzig. Early in the present century, Spearman (161,pp.201-293) improved the methods for calculating correlations and used more factors for the determination of general intelligence. In 1906, Norsworthy (125,pp.25-26) reported on the use of the Seguin form board. Since that time, the form board has undergone many modifications but it has continued to be extremely popular as a nonverbal instrument. In one of the earliest batteries (91, pp.1-53) of mental tests, form boards and picture puzzles were included.

The work of Knox (101,pp.741-747) at Ellis Island led to the development of the Knox Cube Test which was subsequently extended and standardized by Pintner (130,pp. 377-401). Arthur has incorporated this test into her Performance Scale.

In 1914, Healey (90,pp.189-203) presented his Pictorial Completion test. While Arthur (86,p.447) states that scores on this test appear to correlate with language ability more closely than many of the other

performance tests, the pictures reflect the culture of the group in which they originate to a degree that renders them of little value for use in groups of different cultural background (89,pp.425-426).

In 1915, Porteus published the first account of work with his maze test (138,pp.1-194). During the intervening forty years, the test has been widely used in psychological clinics, both by itself and as part of the Arthur Scale. Originally, Porteus intended that the Mazes would supplement the Stanford-Binet scale as a nonverbal measure of "prudence and forethought". Subsequently, there has been a tendency to emphasize the value of the Maze test in revealing planning ability and foresight in dealing with a simple concrete situation. Louttit and Stackman (111,pp.18-25) report correlations between the Porteus and Binet tests ranging from .54 to .69 indicating appreciable communality between the abilities required by these two tests. The correlation coefficients between the Maze test and other well known tests of intelligence have all been positive, suggesting a high degree of saturation with Spearman's 'g' factor. Porteus feels that this reflects the common factor, "planning capacity", that is present in all intelligence tests. This finding is supported by Burt (32,p.277) who, in a factorial study found the Maze test to have a saturation of .716 with the

general factor found in a battery of cognitive tests and a correlation of .667 with general intelligence as assessed by teachers. It is interesting to note that Burt (32,pp.1-467) included the Porteus test in his "Mental and Scholastic Tests" in 1921 and has contributed to its popularity in Great Britain. The author's claim for the validity of his test rests upon relationships between test scores and social ratings. Porteus (139,pp.180-188) contends that one can never arrive at a measure of social intelligence (foresight, planning, and social sufficiency) by validating tests against educational standards. He has published a summary (136,pp.1-219) of twelve studies conducted with mentally defective patients at Vineland, New Jersey. The average Binet correlation with social efficiency was .58 while the average Maze correlation was .68. Porteus (86,pp.539-540) reports that Brundage found the test useful in selecting machine operators. He found a correlation of .498 between maze I.Q. and an industrial efficiency criterion. More recently (139,pp.180-188), he has cited psychosurgical findings of the Columbia-Greystone projects and the New York Brain Study project as supporting the claim that the Maze test is a valid measure of planfulness and that this capacity is closely related to social sufficiency. The claim of Porteus that performance on the Maze test is affected by psychological changes

following prefrontal leucotomy is supported (47,pp.3-41; 48,pp.92-99) but at the same time, Crown (49,pp.49-88) cites evidence that scores on certain verbal tests are similarly affected by leucotomy and that changes in scores on the Maze tests are slight. The Maze Tests were standardized on an age-level basis and the mental-age-score is based upon a success-or-failure performance on the series. In 1942, Porteus (137,pp.1-37) published a report on a method of scoring the quality of performance which is relatively independent of the maze I.Q. This qualitative scoring promises to add to the clinical usefulness of the test. The Mazes have been employed widely in studies of national-racial differences and they are well suited to the task as they are almost completely independent of language, education, or culture and they are universal in their interest and appeal. Porteus has used his Mazes with persons from different cultures and he has published racial norms for his test. These have been trenchantly criticized by anthropologists. Porteus has pointed out that his test is not culturally meaningless to primitive peoples as are other so-called "culture-free" tests since they involve straight lines and simple training. Porteus has performed a service in providing norms for peoples of different cultures, but the meaning of these differences remains to be understood.

During World War I a nonlanguage group test, the Army Beta (186,p.276-283) was developed for use with illiterates. While it proved to be less valid for the purpose than its verbal counterpart (the Army Alpha test) it was sufficiently discriminating to be worth using.

Probably the first series of geometric puzzles to be used in mental measurement were those published by Kent (98,pp.40-50) in 1916.

In 1917, Pintner and Paterson published a combination of fifteen nonverbal tasks to form the first standardized scale of performance tests. The authors employed "chronological age of the child" as the validation criterion (134,p.171). The scale yielded a point score. Percentiles were provided for each age level and tables facilitated the computation of a mental age for each subtest. Manipulative dexterity was involved throughout and in twelve of the fifteen subtests speed was an important factor. Two of the tests involved memory. The scale was overloaded with form boards and hence it was not satisfactory for use with older children. The weight given to some of the subtests was out of line with their discriminative value. Mursell (122,p.134) considers the Pintner-Paterson scale to have been a valuable supplement to highly verbal tests but not a good substitute for them. Arthur (86,p.448) has pointed out the usefulness of the

scale as a handbook for the fifteen nonverbal tests that it included.

Throughout the period of extensive test-development there were many studies designed to evaluate the worth of this performance scale. In 1925, Dashiell and Glenn (53,pp.335-340) compared the intelligence of children of Chapel Hill, North Carolina, as evaluated by the Stanford-Binet and Pintner-Paterson scales, with their economic status. They found that while classifications of the Chapel Hill subjects on their Stanford-Binet scores paralleled the socioeconomic strata, their standings on the Pintner-Paterson Performance scale presented a different picture. The authors have recommended that when studies are being made of group differences in intelligence along lines of socioeconomic cleavage, the use of the Binet scale should be supplemented by performance tests.

In 1920, Kohs (102,pp.357-376) published his first report on his Block Design test. The correlation with Stanford-Binet mental age for 366 cases was .82 $\pm$.01. In the Arthur Point Scale of Performance tests (Form I) the Kohs Block Design test showed the highest discriminative value between successive age groups from 5.5 to 15.5 years inclusive, of any test in the scale except the Seguin form board (86,p.448) Modifications of Koh's test

have been used in several of the nonverbal scales. In the present study, it has been employed as a subtest of the Wechsler-Bellevue Performance Scale. It should be noted here that in 1944 Arthur (15,pp.33-34) devised a stencil design test to replace the Kohs test in the Revised Form II of her Performance Scale in order to avoid the high degree of practice effect shown by the Kohs test in retest scores at higher age levels. Arthur has claimed the Stencil Design test to be a satisfactory measure of general ability at age levels 4.5 to 15.5 years inclusive, and, because of the spatial relationships involved in the problems presented, to be extremely useful in psychiatric clinics.

Yerkes (185,pp.120-292) designed the Army Performance Scale to test foreigners and illiterates. A correlation coefficient of .73 was found to obtain between performance scores and Stanford-Binet mental ages.(185,p.387)

In one of the earliest studies of mental differences among immigrant groups Kimball Young (187,pp.1-103) administered the Army Alpha (verbal) and Beta (nonverbal) examinations to nearly 1,000 twelve-year-olds. The correlation coefficient between these two tests was found to be $.736 \pm .016$. Young found that the combination of Alpha and Beta gave a better diagnosis as measured against outside criteria than either one of the tests, and that the

verbal test was the most significant half of the combined test.

A unique approach to the problem of nonverbal culture-free testing was that of Goodenough (80,pp.1-177) who, in 1926, presented her Draw-a-man test, which had special provision for use with non-English speaking children. For the chronological-age groups, four to ten years, a correlation of .76 with Stanford-Binet ratings is reported. The Goodenough test has frequently been employed in racial comparisons of intellectual ability and five of these studies are cited later in this chapter.

In 1925, Arthur (11,pp.390-416) published a Point Scale of Performance Tests. This was an individual test for use with subjects from five years of age to adulthood, and it has been considered (44,p.478) to be the best of the nonverbal tests. Form I was standardized on a population of nearly 1000 average American school children; approximately 100 at each age level from five to fifteen years. This test and its normative data proved to be satisfactory from the start. Goodenough and Maurer (83,pp. 1-130) have presented evidence of its validity and reliability for comparatively unselected groups. They have cited as correlation coefficients between Arthur I.Q.'s and Binet I.Q.'s values that range from $.54 \pm .08$ to $.81 \pm .04$ at various age levels. In 1943, Arthur (14,pp.1-60),

compared 974 I.Q.'s obtained on Form I with the I.Q.'s obtained by the same subjects on the 1916 Stanford-Binet and found a probable error of less than five points at nearly all age levels. For twenty-five subjects tested on Form I and on the Terman (Form L) the correlation was found to be about the same as that generally reported between nonverbal and verbal scales; r = .73±.09. (86,p.449)

Arthur (12,pp.251-264) demonstrated that dull subjects do not tend to rate any higher on her Performance Scale than on the Binet. She administered Form I to 432 dull subjects who had earned Stanford-Binet I.Q.'s below 95 and found the median algebraic difference between the Binet I.Q. and the Arthur I.Q. to be ± 0. Subsequently, the author (17,pp.276-279) reached identically the same conclusion in a parallel study with 60 feebleminded subjects. At the same time, there has been no general tendency for bright subjects to rate lower on the Arthur performance scale than on the Binet. Arthur (86,p.449) gave both tests to 111 bright subjects all of whom had earned an I.Q. of 115 or higher on the Binet scale and found that the median algebraic difference between the Binet I.Q. and the Arthur I.Q. was +1.0.

Loudon and Arthur (110,pp.599-606) have demonstrated that patients with a reading disability tend to earn a rating on Form I of the Arthur Scale that agrees

more closely with the Stanford-Binet rating obtained after satisfactory reading skills have been developed than does the Binet rating obtained before they have learned to read.

In a study (17,p.279) of the relative difficulty of various tests for the feebleminded, Arthur has shown that such individuals perform no better on nonverbal tests than they do on verbal tests. Of the abilities tapped by the Arthur scale, those most conspicuously retarded were attention span as measured by the Knox cube test, and reasoning ability as rated by the Kohs block design test.

A parallel form (Form II) of the Arthur scale was designed for retest purposes. It had several technical weaknesses and it was inadequately standardized. This form was never extensively used and eventually it was discarded. In 1947, a thorough revision and standardization was completed. The ship test was omitted and the stencil design test was substituted for the Kohs block design test. This revised Form II was constructed to be a reliable alternate instrument for testing the abilities measured with Form I: attention and memory span as measured by the Knox cube test, speed of psychomotor reaction as rated by the Seguin form board, logical thinking and problem solving ability as tapped by the stencil design test, planning ability as evaluated by the Porteus mazes,

and other special abilities as rated by the Healy picture completion test. The test materials are of good quality, convenient and easy to use. From his experience the writer has found the tasks to be intrinsically interesting to children and they appear to stimulate adequate effort. Norms for this revised Form II were derived from the scores of 968 pupils from the same middle-class American district used in standardizing Form I. In the manual Arthur states that the validity is indicated by the "discriminative value" of the test at successive age levels and by the agreement of ratings obtained from the revised Form II with those obtained from Form I and with those obtained from the Binet scale. The revised Form II shows a correlation of .78 with Binet I.Q.'s for 171 subjects (86,p.450). Arthur submits as evidence of the reliability of the revised Form II, the median differences between I.Q.'s secured by it and those secured by Form I or by the Binet test (16,p.23). The writer is of the opinion that the data presented in the manual as evidence of the validity and reliability of this test are hardly adequate for a scale that has such widespread clinical use.

Following the publication of the first Arthur test in America, the next noteworthy development was the appearance of the Alexander Performance Scale (31,p.435)

in Scotland. This scale consists of three tests, the Passalong test (which is a manipulative puzzle standardized as a test for the upper age levels), a shortened version of the Kohs Block Design test and the Cube Construction test. Apart from this scale, these subtests have demonstrated their usefulness in other contexts. Thomson (169,pp.1-58) has demonstrated that the Cube Construction test, the Kohs Block test and the Healy picture completion test, gave the best multiple prediction of the Binet I.Q. of Scottish children and that the addition of other performance tests did not materially improve the predictive efficiency of the battery.

While the Minnesota Pre-School Scale (82,pp.1-44) was designed for use with subjects much younger than those in the present study, its development is noteworthy in that it was the first performance scale to be divided into language scores and nonlanguage scores. In 1942, Goodenough and Maurer (83,pp.1-130) reported on the predictive value of the Minnesota Pre-School scale. They found that the group nonverbal items were more closely related to all later tests than were many of the verbal items. If the general-factor hypothesis is accepted, these nonverbal tests appeared to be a better measure of the "g" factor at pre-school ages than are many of the verbal tests.

Subsequently other performance scales appeared. Typical of these was the Cornell-Coxe (46,pp.1-88) which included seven subtests most of which had appeared previously in the Army Performance Scale. While the authors reported a correlation of +.79 between Stanford-Binet mental ages and the Cornell-Coxe ratings (46,p.32), it would seem to the writer that the size of the standardization group (306 subjects) was hardly adequate for an age range extending from kindergarten to the high school level.

Simultaneous with the publication of the new scales, research was being conducted into the effectiveness of existing instruments. In 1936, Weisenburg, Roe and McBride (183,pp.1-155) administered a wide variety of nonlanguage tests to seventy adults, carefully selected to represent the middle levels of the population. They found that most of the subtests on the Pintner-Paterson Performance Scale were too easy for average adults. While the mare-and-foal-test, Seguin form board, and the substitution tests alone were found to be moderately discriminative for adults, none of these was expecially interesting to them or particularly valuable in the study of normal adult intelligence. Of the Pintner nonlanguage scale only the reversed-drawing subtest presented difficulty to adults. The Goodenough drawing test

distributed well but the scores were not comparable with similar scores made by children. The Porteus Maze was well received and appeared to hold the interest of adults as much as that of children.

The Leiter International Performance Scale (103, pp.1-95) is noteworthy in that the standardization and location of items in the scale followed the procedure used by Terman in the Revised Stanford-Binet scale. For Terman, the amount of information a child acquired through incidental learning was an index of brightness. On the other hand, Leiter held that the ability to cope with new situations was a truer indication of intelligence. Accordingly, he arranged 68 items in order of increasing difficulty, four at each age from the 2-year to the 18-year level. Instructions were given in pantomime and the test employed a simple technique of matching colors, objects, relationships, etc. Because the situations were novel and the material unique, the possible effects of practice or coaching were minimized. While the test has established itself as a useful diagnostic instrument in clinics, it has not come to be used widely in school situations as a measure of general intelligence. Bessent (22,p.234) has published a note on the validity of the Leiter scale. In a study of twenty cases, he found a correlation of $.92 \pm 0.035$ between Binet and Leiter I.Q.'s.

In a study closely related to the present one, Tate (168,pp.497-501) attempted to evaluate the relative freedom from cultural influence of the Leiter, the Arthur and the Stanford-Binet Form L. These tests were administered to 108 subjects that were chosen in equal number to represent four distinctly different socioeconomic groups. All three tests differentiated significantly between all pairs of the experimental groups except between the two professional groups. The Leiter appeared to be no more "culture-free" than the Arthur or Binet. The mean scores for all groups on the Leiter scale were consistently lower than those of the Arthur or the Binet which were approximately equal for any given group. Tate found that the Leiter scale correlated as highly with the Arthur and the Binet as those tests did with each other. She has pointed out that notwithstanding the recent restandardization by Arthur, the Leiter is in serious need of a revision of published norms.

Cultural difficulties commonly found in verbal tests may also appear in those performance tests that require pictorial interpretation and manipulative skills. In order to minimize these difficulties perceptual tests such as Raven's Progressive Matrices and Cattell's Culture-free tests are now commonly used and the instructions are pantomimed. These perceptual tests have an

interesting origin (38,p.167). Davey (55,pp.27-48) showed that pictorial tests of intelligence involved the same "g" factor as other intelligence tests in current use in Britain. Line (108,pp.1-148) discovered that a test involving the eduction of relations between simple geometrical shapes was highly saturated with the general intelligence factor. Stephenson (166,pp.334-350) confirmed that the same "g" factor ran through verbal and nonverbal tests alike. Lorge and Arsenian (109,pp.520-522) (9,pp. 287-301) demonstrated that the Spearman Visual Perception test showed significant differences between racial groups in situations in which traditional tests would have given ambiguous results. Cattell (38,p.168) has referred to a large scale factor analysis at Mooseheart, Illinois, (Spearman-Holzinger Unitary Trait Study, University of Chicago, 1935) that showed the same perceptual test to be highly saturated with the "g" factor. Penrose and Raven (128,pp.97-104) designed a new series of tests to avoid the disadvantages of perceptual tests in use at that time. Their premise was that the general intellectual ability of an individual could be defined by the complexity of the relations which he is capable of handling.

Raven's Progressive Matrices appeared in 1938 providing a nonlanguage perceptual test for measuring

intelligence. The author has described it as

> "a test of a person's capacity at the time of the test to apprehend meaningless figures presented for his observation, see the relations between them, conceive the nature of the figure completing each system of relations presented, and, by so doing, develop a systematic method of reasoning." (147,p.1)

It has been suggested by Raven that this scale should be supplemented by a vocabulary test. From the results of an experimental survey carried out in Colchester (144,pp.16-34) a series of sixty matrix tests was prepared for general use.

While the author has specified in his manual (147, p.2) that the Progressive Matrices is not intended by itself to be a test of general intelligence, he has defined the five grades assigned by the test in terms of general intellectual capacity (147,p.9). Slater (160,pp. 20-21) has claimed that Raven's test ranks second only to Binet's as a test of general intelligence. Westby (31, pp.418-422) has reported that factor analysis in the British Services suggests that the Progressive Matrices test is an almost pure "g" test with a small loading of some spatial perceptual factor and that the latest data for its reliability agree with the figure of 0.88 which testing in the British Services revealed. The manual (147,p.2) has cited from studies by Burt a correlation coefficient of

0.86 with the Terman-Binet test and a "g" saturation of 0.82. The high "g" saturation was confirmed by Adkins and Lyerly (1,pp.1-122). In 1949, Raven (146,pp.12-19) found a correlation of .855 between Matrix Score and Terman mental age, and when combined with his Mill Hill vocabulary score this coefficient was .918. The norms for the individual test have been based on the scores of 660 subjects selected by random sampling in the county borough of Ipswich. The group test norms have been established on the results of scores of 1407 children (randomly selected) and 3665 adult males. Raven's case notes (145,pp. 137-150) during the standardization have shown that verbal fluency sometimes influenced Binet I.Q.'s while not affecting Matrix test scores. Thus, when used with defectives, the Matrices has been able to differentiate backwardness due to specific defects in reading, speech or education from genuine intellectual defect. Apart from suggestions for a better grading of the items, there has been little indication of need or desire for a revision of the 1938 scale. However, the author has developed two derivatives, Colored Progressive Matrices (1947) Sets A, Ab, B and Progressive Matrices (1947) Sets I, II. While the first has been constructed to disperse scores and to discriminate further among a group within the lowest quarter of the population, the second is a similar and successful

(66,pp.104-110) attempt to deal with the top quarter.

For comparative studies, the scale has been used internationally: in Britain and continental Europe, in America (36,pp.233-241), with Zulus of South Africa (126, pp.68-70), in Israel (3,pp.156-159) and in South America (149,pp.347-352) (150,pp.81-114) (151,pp.1-25). While Alou-Bakaliar (3,pp.156-159) found that Yemenite children scored lower than occidental children in Tel Aviv, and Notcutt (126,p.68) found that the performance of Zulu children was inferior to that of Raven's standardization group, most comparisons have agreed with Rimoldi's (149, p.351) finding that there is a striking similarity between results obtained in different populations, in different countries and under different testing situations.

Cassel (36,pp.233-241) has attempted to arrive at a qualitative evaluation of performance on the Progressive Matrices by studying the patterns of incorrect responses made by mental defectives at Vineland, New Jersey.

Raven (146,pp.12-19) has administered both the Matrices and the Mill Hill Vocabulary Scale to 8500 subjects ranging in age from four to 65 years, and, with this data, he has attempted to trace the normal changes, as age advances, in a person's capacity to reason by analogy and to recall information. From these results, it has been possible to calculate percentile norms for each test,

for children and for adults up to the age of 65.

Several studies (97,pp.140-150) (66,pp.104-110) (160,pp.20-21) (158,pp.238-239) have related to the validity and the reliability of the Progressive Matrices test. Keir's London study (97,p.149) evaluated coefficients of validity (.56) and reliability (.70) that were significantly lower than those found by Raven or Burt. Similarly, Sinha (158,p.238) found a low validity coefficient (0.54) when the Matrices scores were compared to those on the Revised Stanford-Binet and Simplex Junior intelligence test. Besides raising questions about the reliability and efficiency of the test as a whole, Keir, Sinha and Horton have indicated how the arrangement of items might be improved. Levine and Iscoe (104,p.10) found a correlation of .55 between Matrices scores and total score on a shortened form of the Wechsler-Bellevue scale. The Progressive Matrices correlated more highly with the block design test (.63) than it did with vocabulary (.48) or comprehension (.21) tests. Martin and Wiechers (115,p.144) found a correlation of .91 between Colored Progressive Matrices Scores and Wechsler's Intelligence Scale for Children. The intercorrelations among Matrices and various subtests ranged from .74 on the block designs to .47 on the information test. Stacey and Carleton (162,pp.84-85) gave the Colored Matrices, the Revised Stanford-Binet (Form L) and the

Wechsler Intelligence Scale for Children to 150 subjects who had been referred to a State school as possible mental defectives. The correlation between Matrices scores and Binet I.Q.'s was .71 and that between Matrices scores and Wechsler total weighted scores was .62. Correlations of the Matrices scores with weighted scores on the subtests ranged from .48 on the Picture Completion test to .28 on Coding and Mazes. This experiment was repeated (163,pp. 86-87) with 172 subnormal adult subjects and the children's form of the Wechsler was replaced by the adult test. The correlation between Matrices scores and Binet I.Q.'s was .86 and that between Matrices scores and Wechsler full scale I.Q's was .68. Correlations of matrices scores with weighted scores on the subtests ranged from .60 on block designs to .29 on arithmetic. Green and Ewert (84,pp. 139-142) administered by slides the Colored Matrices to 1213 school children in Rochester, Minnesota. Besides providing a wealth of normative data, the authors reported a correlation between Matrices scores and Otis mental ages of .78. Because scores on the Progressive Matrices correlated with the more verbal intelligence tests to about the same degree that they did with nonverbal tests, they concluded that the Matrices should not be thought of as a test of nonverbal reasoning ability but rather as a test of "fairly complex reasoning processes". Bolton

(24,pp.629-633) reported a correlation of .80 between Arthur mental age scores and scores on the Matrices for 33 non-English speaking immigrant children. Subsequently, he administered the Matrices to 1160 fourth-graders. He found correlations between Matrices and Pintner Nonlanguage scores, of .53 and .51; between Matrices and Henmon Nelson scores, .47; between Matrices and Terman-McNemar scores, .58; between Matrices and Otis Gamma scores, .40. He found a corrected split-half reliability coefficient of .90. Bolton claimed that it was possible to derive usable I.Q.'s from Raven's raw scores, but that these have little value in predicting scholastic achievement (subject marks) at the fourth grade level. Desai (59,p. 60) administered the Progressive Matrices and the Wechsler-Bellevue verbal scale to 190 male subjects in a mental hospital. He found a correlation between scores on these tests of .573 (or .648 when the coefficient is corrected for attenuation). In a recent study, Levine and Iscoe (105,pp.307-308) administered the Progressive Matrices, the Metropolitan Achievement Tests, the Chicago Nonverbal and the Wechsler-Bellevue Performance tests to 73 adolescents in the Texas School for the Deaf. The correlation between Matrices scores and Metropolitan achievement scores was .423, that between Matrices and Chicago nonverbal tests was .413, and the correlation

between the Matrices and Wechsler Performance scores was .552. Although this last coefficient is far from being sufficiently high for accurate individual prediction, it would seem to warrant continued use of the Progressive Matrices with deaf subjects, particularly in those cases where the hearing loss is complicated by other factors that hinder or prevent the use of performance tests.

The Wechsler-Bellevue Scale (182,pp.1-258) was developed in a mental hospital primarily for the diagnosis of mental impairment in adolescents and adults. From the time of its publication in 1939, the scale was popular among psychometrists; it was an individual test, it provided both a verbal and a performance scale, and it could be used with subjects from ten to seventy years of age. Over the years test-users have found this well-conceived instrument to be an adaptable tool for use in clinic, school, guidance bureau or research center. Current studies seldom question the validity of this test as a measure of intelligence (142,pp.410-422). In fact, the vast popularity and wide usage of the Wechsler-Bellevue have made it a commonly used basis for comparison and validation of newer instruments and more recent techniques (68,pp.268-269).

According to Wechsler, the test presents the opportunity of obtaining a measure of the individual's

global intelligence along with the configuration of those elements that compose this global entity (182,p.3). Clinical psychologists (67,pp.71-85) (113,pp.217-229) have been enthusiastic about the diagnostic potentialities in its differential yet homogeneous composition and to this end there has been a continuous flow of "pattern" studies that have not proven to be particularly rewarding. Typical of these are the current investigations of Matarazzo and associates (116,pp.201-205) (117,pp.131-134) (118,p.218) (33,pp.280-282) into the relationship between manifest anxiety and performance on Wechsler subtests.

In discussing the validity of his instrument, Wechsler has furnished evidence of agreement between Wechsler-Bellevue I.Q.'s and the scores on several other tests (182,p.134), with teacher ratings (182,p.130), and case study data (182,pp.130-132), but he has indicated that a better claim for the test's validity might rest on the fact that it has given satisfactory service in clinical practice (182,p.127). He has cited a study employing test-retest at varying intervals with 52 subjects that yielded a reliability coefficient of .94 for the full scale (182,p.133). Derner, Aborn and Canter (58,pp.172-179) have summarized studies of the test-retest reliability of the Wechsler Scale and they have reported average coefficients for 158 normal individuals, tested at well-

controlled intervals, of .90 for the full scale, and from .62 to .88 for the subtest. Goldfarb (78,pp.503-507), Rabin (142,pp.410-422), Watson (181,pp.61-68), and Rabin and Guertin (143,pp.211-248) have summarized studies of the Wechsler-Bellevue and these revealed trends for the Wechsler-Bellevue and Revised Stanford-Binet scores to correlate from .78 to .93 with heterogeneous groups in age or mental ability, and about .62 when the groups are more homogeneous. The verbal scale correlated more highly with the Stanford-Binet and other traditional tests of intelligence and achievement than did the performance scale. Wechsler reported the correlation (corrected for attenuation) between performance and verbal I.Q.'s to be .83±.018 (182,p.133).

Sartain (156,pp.237-239) found that the correlations of the Wechsler-Bellevue, the Revised Alpha Examination, the Otis Self-Administering Test of Mental Ability, the American Council on Education Psychological Examination and the Revised Stanford-Binet Scale with grade-point-averages of fifty college freshmen failed to reveal significant differences among the tests. The Wechsler scores correlated .740 with the Alpha, .697 with the Otis, .774 with the Stanford-Binet, .692 with the A.C.E. Psychological Examination scores, and .534 with grade-point-average. Altus (4,pp.42-44) found a validity coefficient

of .579 between Wechsler scores and an acceptance-rejection criterion among recruits at an Army special training center.

There have been many attempts to develop abbreviated forms of the Wechsler-Bellevue scale (75,pp.101-108) (141,pp.320-324). Mech (120,pp.241-260) has made an item-analysis of seven subtests. He found that the Block Design test had the highest discriminative value of all performance subtests. Other studies such as that of Burik (29,pp.33-42), have investigated the role of motor factors in some of the performance tests.

The Wechsler-Bellevue has been an effective instrument for racial comparisons of mental ability. Davidson, et al (56,pp.490-491), found that among a sample population of psychoneurotic patients Negroes showed a significantly lower score than Whites on the Arithmetic subtest and on all subtests of the performance scale. The authors have suggested that the time element in performance tests handicaps Negroes whose performance of psychomotor perceptual functions is slower than that of whites and who have little incentive to do things rapidly.

Glaser (77,p.241) investigated the intelligence of Jewish immigrants and pointed out the cultural difficulties encountered with the Information, Digit Span, Digit Symbol, and Picture tests (143,p.221).

The movement to develop culture-free tests has gained impetus from the work of Cattell (37,pp.114-131) (41,pp.1-411) who has persistently objected to the continued use of Binet-type scales.

The Cattell test (38,pp.161-179) (43,pp.81-100) was an attempt to provide a measure of general mental ability free from both verbal education and the acquired skills tested by many of the nonverbal performance tests. As originally published, the Cattell test consisted of 159 items in seven untimed subtests including series, mazes, classifications, pool reflections and matrices. Subsequently, four item analyses were made and only 72 items survived. The maze tests were deleted because they were low in discriminative value. Time-limits were imposed, although the limits were long enough to enable almost everyone to finish. The range of difficulty of the items suited subjects from those with a mental age of twelve years to superior adults. The advantages of the test have been the convenience of its materials, the possibility of its use as a group test, and the ease of scoring. On the other hand, the possible monotony of the test when successive tasks all depend on visual discrimination of differences and the lack of an opportunity to observe the response of the subject in different situations have been its main disadvantages. The author cites correlations

with the Army Alpha from .53 to .60 for four groups of high school students (39,p.4). A reliability coefficient based on the scores of 121 students is given as .88 (39, p.3).

In a series of three carefully designed experiments using the Terman-Merrill revision of the Binet, the arithmetic sections of the American Council Psychological Examination, the Arthur Performance Scale, the Ferguson Form boards and the Cattell test, Cattell, Feingold, and Sarason (43,pp.81-100), demonstrated that while the Terman-Merrill is about as valid as the Cattell test, and others (notably, the Arthur) are quite as free from susceptibility to cultural influence, the Cattell test is the only one which combines both of these advantages. Noteworthy is the design of their investigation of the effect of acculturation. A group of adult immigrants to America and a control group of native Americans of the same age and I.Q. were given the five tests and then were retested after an interval of eleven weeks. Improvements on the test were expressed in terms of the mean standard deviation of the control group. The authors reasoned that since none of the gain in the retest-scores of the control group could be due to cultural factors, then any such improvement must be attributed entirely to test sophistication and practice. The subtraction of this value from the

gain of the immigrants should yield the gain due to acculturation. The Terman-Merrill and the A.C.E. tests appeared to be much more susceptible to cultural influences than the Cattell or the performance tests. The same authors have evaluated the mean intercorrelation of scores on each test with the pooled score for natives and immigrants. They found an overall-mean coefficient of .52 for the Cattell, .50 for the Terman-Merrill, .44 for the Arthur, .40 for the A.C.E., and .18 for the Ferguson form boards (43,p.97). Further, they computed a consistency coefficient (corrected for length of test by the Spearman-Brown formula) of .90 for the Cattell test (43,p.95).

Drake (31,pp.384-385) has reported investigations in which the Cattell test correlated .84 with the Wechsler-Bellevue Intelligence Scale, .83 with the Army General Classification Test and .84 with the Otis Self-Administering Test of Mental Ability.

Tilton (173,pp.17-19) has made a survey of the validity, reliability and usefulness of the Cattell test. He administered the test to 75 high school seniors in Kent, Ohio, and found that the Culture-free test does not correlate as well (.36) with high school grades as does the Henmon-Nelson (.80) or the Otis (.62), nor as well (.36) with teachers' ratings of intelligence as does the Henmon-Nelson (.71) or the Otis (.71). The author has accounted

for the difference in terms of the absence of the memory factor in the Cattell and its presence in verbal tests and the usual overemphasis in school on verbal and memory factors as represented by school grades and teacher ratings of intelligence. In the same study Tilton presents evidence that scores on the Cattell test correlate more highly (.84) with Wechsler-Bellevue than with the Otis (.66) or Henmon-Nelson (.60) scores.

Vernon (178,pp.237-244) has shown that a culture-free test is no less subject than other tests to "test-sophistication".

Pierce-Jones and Tyler (129,pp.109-114) have used the American Council Psychological Examination and the Cattell Culture-free test to see if two groups of psychology students dissimilar in motives and in certain types of experiences (one group from an arts and crafts college and the other from a college of education) will differ significantly on Cattell's Culture Free Test. The sexes were segregated and groups from the two schools were matched on the basis of percentile rank on the 'Q', 'L' and 'T' scores of the A.C.E. Psychological Examination. Four of the six differences in the mean Culture-free scores were significant at the five per cent level or better and in all cases the difference was in favor of the arts and crafts college students suggesting to the authors that the scores

on the Culture-free test were to some extent related to experiential factors and interests. Since one purpose in administering intelligence tests is to predict school achievement this study was concerned also with the relations between ability as measured by each of these tests and performance in psychology courses. The Cattell test proved to be a poorer predictor of scores on two psychology examinations than were the 'Q', 'L', or 'T' scores of the A.C.E. psychological examination. The Cattell test correlated with academic success to the same extent in the two groups; and the authors have suggested that the attempt to hold a cultural factor constant has reduced correlations to a lower level. They have concluded that there appears to be no particular point in preparing a culture-free test which does not differentiate between various groups and at the same time does not materially predict academic success. (129,p.113)

Cattell's IPAT Test of g: Culture-Free (1949) (42,pp.1-9) has several refinements and improvements over his original test. It is shorter, more convenient to give, and it is available in two equivalent forms at three different levels. In his manual (42,p2) Cattell has referred to research (43,pp.81-100) which has shown that immigrant groups do not have the large difference between first and later testings on the culture-free tests that is shown on

the A.C.E. or Binet instruments, and that the "g" saturation (as measured by a pool of standard intelligence tests) is as good as that of verbal tests and much better than that of performance scales. The author has cited "g" saturations for this test ranging from .53 to .99 in various experiments (42,p.2) and that its reliability corrected for full length has been .70, .86, .87, and .92 with various groups of children. Drake (31,p.401) found for scale 2 a correlation of .56 with the Revised Stanford-Binet scale and .36 with the Stanford Achievement test; this value being 11 points higher than the Stanford-Binet "r" of .25 with the Stanford Achievement tests. Drake cited another study in which for ten different groups of high school and college students scale 3 correlated on the average more highly (.51) with grade point averages than did the Otis Self-Administering Tests of Mental ability (.35) and the intercorrelation of these two tests was .73. For a group of 32 college students, the Cattell correlated .59 with the American Council on Education Psychological Examination.

Cattell (40,p.157) has provided evidence that as intelligence tests have their scholastic contamination reduced advancing to more culture-free forms, the standard deviation of I.Q.'s increases from twelve to twenty-four or twenty-five points. He has pointed out the

desirability of providing with any intelligence test both a "classical" I.Q. standardization and a table of standard-scores in which I.Q.'s are reduced to some agreed standard deviation. Cattell has carried out these recommendations with respect to the IPAT test adopting a sigma of 24 points for the standard score I.Q.

Vernon and Parry (179,pp.1-324) have reported on the procedures used and the results obtained in selecting personnel during the second World War for the Royal Navy, the Imperial Army, the Auxiliary Territorial Service, and the Royal Air Force. They found (179,pp.188-191) that the chief differences in outcome of standardization of perceptual tests (as opposed to the traditional verbal tests) were that the curve of increase of score with age was likely to flatten out a bit earlier and that the standard deviation of I.Q.'s was likely to be significantly larger.

Considerable impetus has recently been given to the movement for culture-free tests by the research of Eells, Havighurst, Davis (63,pp.1-388) and their associates at the University of Chicago. They administered the Henmon Nelson, Kuhlmann-Anderson, and the Otis Alpha (both verbal and nonverbal) to 2200 nine-and ten-year old pupils and the Terman McNemar, Otis Beta, Thurstone Spatial Relations and Reasoning tests and the California

Test of Mental Maturity to 2400 thirteen-and fourteen-year olds. The authors have claimed that children from high socioeconomic backgrounds have an advantage over children in low socioeconomic circumstances. Herrick has written,

> "The research is close to unanimous in showing that there are significant differences in intelligence test performance of children and youth from different socioeconomic backgrounds, with children from the higher levels always securing the higher intelligence test scores." (63,p.12)

High status children showed superior performance on verbal items particularly. The difference between high and low status was least on picture, geometric designs and stylized drawing items. The authors have argued that culture-fair tests should avoid all types of problems on which any one socioeconomic group has had more experience or practice than another. To this end, they have chosen basic mental problems for new tests of intelligence, (viz., the Davis-Hess Individual Test of Intelligence and the Davis-Eells Games) and they have tried to express them in symbols and situations common to all occupational groups. The publication of these tests and the report of the cooperative study at the University of Chicago have provoked widespread reaction from makers and users of tests. Stenquist and Lorge (165,pp.184-193) have dealt

with the implications of their findings. McNemar (112, pp.370-371) and Tyler (177,pp.288-295) have criticized the statistical methods employed. Tyler has pointed out that when consideration is given to factors such as the reliability of the tests, the correction of correlations for attenuation, the inequality of I.Q. units within each test and from test to test, and the difference in relative difficulty of the tests, the data give little positive support for the conclusions that are drawn. He has demonstrated that if the I.Q. scores had been converted to standard scores the difference between the regression lines for verbal and nonverbal intelligence against status would largely disappear (177,p.293). Budd (28,pp.333-334) has pointed out what he believes to be the reason that educators will not be impressed with culture-fair tests.

Rosenblum, Keller and Papania (152,pp.51-54) have evaluated the performance on "Games" of a group of mentally handicapped boys of lower social class standing and the results were compared with those obtained on the California Test of Mental Maturity, the Revised Stanford-Binet, and the Wechsler Intelligence Scale for Children. For this group of subjects, the mean score on Games was not significantly higher than those on the other measures. The Davis-Eells's test failed to reveal intellectual potential not tapped by other tests that are presumed to

be culturally biased. The authors have suggested, however, that this may be due to the subjects' deficiency in problem-solving-ability and their lack of ability to abstract. Whatever may be the explanation, this study has shown that performance on Games is as much affected by loss in ability to do abstract thinking as are other intelligence tests.

Allen and Besell (2,pp.394-395) have evaluated the intercorrelations among group verbal and nonverbal tests of intelligence. They found that the intercorrelations among the Otis, Henmon-Nelson and Modified Alpha ranged from .66 to .73 but the correlations of these tests with the Chicago nonverbal examination were .39, .31 and .31, respectively. The low correlations between nonverbal and verbal tests suggest that these tests may be tapping different functions and the authors have recommended the inclusion of both verbal and nonverbal items in a well-rounded testing situation.

Investigations such as Thurstone's extensive study of Primary Mental Abilities (171,pp.1-121) (172,pp.1-94) and Tilton's factor analysis (174,pp.169-179) have tended to reassure psychologists experimenting with performance scales that a wide variety of abilities can be measured without having recourse to verbal materials. Thurstone showed that space perception ability is an important

consideration in measuring nonlanguage intelligence. Fils (65,pp.113-119) found that the correlations between two tests of space perception and a test of nonlanguage intelligence (California Test of Mental Maturity) were: .34 and .43. The authors have concluded from the low correlations that the three tests were measuring different factors. At the same time they have pointed out that they are sufficiently high to suggest that the nonlanguage test includes spatial relations ability as a significant item.

Newland and Lawrence (124,pp.44-47) administered the Chicago nonverbal examination to an east Tennessee Negro population and found that these Negro children scored not less than two years lower than the standardization sample. The authors showed, too, that for this population the test was inadequately discriminative.

Jones, Hey and Wall (94,pp.160-172) modified the Kohs Block design and the Cube Construction tests so that they could be given as a group test. When the two subtests were weighted equally, the group performance scale gave a prediction of a complex criterion of two verbal and two nonverbal tests weighted equally of 0.698. This rose to .732 when the Kohs and Cube tests were weighted 3:1, respectively. The reliability of the scale was 0.850, which is as good, if not better than that yielded by most individual performance tests (94,p.165).

Related investigations into the effectiveness of measures of intelligence have been numerous and varied. Psychometrists, today, are more cautious in the use of traditional tests than they have ever been in the past, and test-makers are paying respectful attention to matters of validity, consistency, and ability to discriminate which are basic considerations in test construction.

Racial Comparisons of Mental Ability

There have been numerous attempts at racial comparisons of mental ability. While many of these investigations have been designed to serve a purpose different from that of the present study, most of them have been concerned with the intelligence testing of subjects of different national groups. Further, in some instances tests identical with those used in this study were administered.

The earliest comparisons of the achievements of different races were made largely on a subjective basis. Noteworthy among them was Galton's scale (70,pp.316-350) for estimating the worth of different races by comparing eminent men in each national group.

Early in the present century, objective measurement was introduced into comparative studies. At the World's Fair of 1904, Woodsworth (184,pp.171-186) and

Bruner (27,pp.1-113) applied a few simple tests of sensory acuity to primitive subjects. They found that the keenness of the senses was about on a par in the various races and that the primitive groups did no better than the norms for white people. The performance of Negritos and Pygmies was inferior to that of other groups on a form board test (184,p.181). Woodsworth suggests that this result indicates the intellectual inferiority of these groups if the test is a fair one.

With the development in 1916 of the Stanford-Binet and the subsequent popularity of mental testing, researchers were no longer content with the simple sensori-motor tests. Psychologists seeking to compare the intellectual characteristics of national groups were quick to employ the new tests. By the mid-twenties, scores of investigations had been made. Some of these included German immigrants as subjects. In a study of retardation in the schools of several cities in northern Michigan, Brown (26, pp.324-327) administered the Stanford-Binet to 1700 children of immigrants. He found the following median I.Q.'s for the national groups: Norwegian, 103.75; German, 102.3; Swedish, 101.9; English, 101.75; Austrian, 99.5; French, 95.4; Finnish, 90.0; Slovak, 85.6; Italian, 77.5. All Germanic groups tested higher than any of the non-Germanic groups. Davenport and Crayton (54,pp.127-134)

tested 102 immigrants and found Germans highest in leadership and pertinacity. In summarizing these studies, Pintner (131,pp.292-295) (132,pp.1-555) points out that immigrants from northwestern Europe gave good accounts of themselves on intelligence tests (132,pp.354-355).

Typical of the many studies of racial mental differences are the extensive investigations of Feingold and Hirsch. Feingold (64,pp.65-83) administered an adaptation of the Army Alpha test to 2,353 college students who were children of immigrants and published the relative standings of the national groups. Hirsch (72,pp.239-240) administered the Pintner-Cunningham; Dearborn A and C tests (which are largely nonlanguage tests) to 5500 subjects representing 16 races and compared the median I.Q.'s of the ethnic groups.

In a study of Hawaiian groups, Porteus (135,pp. 57-74) found that in the Stanford-Binet test, Anglo-saxons ranked first, Chinese second, Japanese third and Portuguese last. On the Porteus maze tests, the Japanese boys were superior to all other race groups up to the age of ten, and beyond ten the Anglo-saxons were superior to other race groups.

More recently, Mann (114,pp.366-395) has considered the suitability of the Binet, Goodenough, Porteus, Healy and Leiter tests for the measurement of race

differences. He has pointed out the intrinsic difficulty of the problem, the inadequacy of the methods and the instruments employed, and the emotional bias attached to the concept of white race superiority. (114,p.366). He has concluded that until new and valid techniques are established, the problem of race differences among primitive peoples will remain unsolved (114,p.391).

Over the years, students of racial psychology have made numerous comparative studies of the intelligence of native Indians with that of whites in American society. Rowe and Rowe (72,p.75) administered the Binet test to 268 Indians and found 94 per cent of them to be below the norms for whites on the basis of chronological age. Subsequently, Hunter and Sommermeier (92,pp.257-277) tested 715 mixed and full-blooded Indians and found a correlation of 0.41 between degree of white blood and I.Q. (92,p.263) This finding was substantiated by Garth (71,pp.382-389) who administered the National Intelligence Test to Indians of various tribes and localities as well as to Mexicans and other ethnic groups. He found that the Mexicans performed better than full-blooded Indians but not so well as mixed-blood Indians. One spectacular finding by Garth was that for Indian children there was a rise in I.Q. with school grade (71,p.388). This caused him to weight the factor of schooling in test performance.

Haught (87,pp.137-142) administered the Pintner-Cunningham, the National, and the Terman group test to children in the appropriate age ranges and concluded that Indians made lower scores than whites because they were lower in native-ability (87,p.142).

Comparative investigations such as these have been closely studied and severely criticized by anthropologists who would attribute differences in test performance to cultural differences rather than to innate differences in intelligence. Blackwood (23,pp.1-120) has stressed the importance of language, motivation, selection and social-status. Klineberg has emphasized the importance of the speed factor (100,pp.159-160). He administered (99,pp.1-111) the Pintner-Paterson series of tests and found that Indian children take longer with form boards but make fewer errors than do white children. There was no significant difference in the total number of points obtained on the Pintner-Patterson Point scale between the Indian and white group because the Indians made up in accuracy for their inferior speed. It is interesting to note that Klineberg found no correspondence of high score on this scale with degree of white blood (99,p.107).

Much has been written about the influence of socioeconomic status on intelligence-test performance. An early attempt (8,pp.179-183) was made by Arlitt to

determine the relative influence of race and social status. This study demonstrated clearly that race norms must take into account the social status factor. Neff (123,pp.727-757) pointed out that children of the lowest socioeconomic classes score on the average about twenty points below children of the professional classes on the Stanford-Binet test (123,p.729). Neff has protested against the use of standard tests for measuring the capacity of individuals from different social levels within our society (123,p.754). In 1940, Skeels (159,pp.281-308) summarized the Iowa studies and pointed out that intelligence was responsive to environmental changes. He quoted a mean Binet I.Q. of 89.7 for 42 five-year-olds in the Iowa Soldiers' Orphan Home as evidence that institutional residence tends to depress the intellectual development of children (159,pp.284-286). Reference has already been made to the contentions of Eells, Havighurst, Davis (63, pp.1-388) that performance on traditional-type intelligence tests is largely dependent on socioeconomic status and cultural background. Gellerman and Hays (76,pp.177-179) have proposed a correction for the confounded effects of cultural variation in intelligence quotients. Thorndike (170,pp.321-338) has discussed the possibility of using community factors as predictors of intelligence.

The influence of language deficiency on test-performance was effectively demonstrated by Jamieson and Sandiford (93,pp.536-551) who administered tests to 717 pupils in Indian schools of Ontario. The median I.Q.'s obtained were as follows: National Intelligence Test, Scale A, Form I, 80; Pintner Nonlanguage Test, 97; Pintner-Paterson Scale of Performance Tests, 96; Pintner-Cunningham Primary Mental Test, 78 (93,p.548). When it is realized that the first of these is predominantly a verbal test and that the last requires detailed instructions given orally in English, the severe language handicap of Indian children on verbal tests becomes apparent. In the same study, the authors found that the performance of pupils in day schools surpassed that of the institutional pupils of residential schools (93,p.548). Probably, of greater significance was their finding that monoglot Indian pupils surpassed the bilinguals in all tests except the performance scale on which the bilingual children obtained a slightly higher median I.Q. than the monoglots (93,p.549). The authors conclude that the verbal nature of the National and Pintner-Cunningham tests contributed to the poorer showing of the bilinguals.

Garth and Smith (74,pp.376-381) found that for Indian children I.Q.'s obtained on a performance scale were ten to fourteen points higher than those obtained on

a verbal test. These children consistently showed a performance on the Pintner-Paterson scale more nearly equal to white performance than they did on the verbal test.

Arthur (13,pp.188-195) administered the Arthur Scale of Performance Tests and the Stanford-Binet Scale to Indian children and found the median I.Q. to be considerably higher on the Arthur test (13,pp.191-193). It should be noted that the Binet results for many of the subjects were invalidated not only by limited environmental experience, but also by a lack of understanding of English as most of these children heard English spoken only in school. The author has suggested that the revised point scale of performance tests is probably better suited for the testing of Indian children than the Binet because it allows adequate time for change of mental set. There is no jumping from task to task as in the case of the Binet. The Arthur scale does reward speed but not haste.

Havighurst and Hilkevitch (89,pp.419-433) administered the Arthur Point Scale of Performance tests to 800 Indian children and found that they did about as well as white children. The earlier finding of Klineberg (99, p.107) that Indian children work more slowly than white children was contradicted (89,p.431) by the results of

this study. The authors observe that although the Healy picture completion tests can be given without the use of verbal instructions, they reflect the culture of the group in which they originate to a degree that would invalidate their use in groups that differ widely in cultural background from the standardization group (89,pp. 425-426).

Turner and Penfold (176,pp.31-44) have recently completed an extensive investigation of the scholastic aptitude of 240 Indian children on the Caradoc reserve in Ontario. The authors established a control population of 215 white children from the surrounding rural districts. To both groups were administered the Otis Quick-Scoring Test (Alpha Form A, nonverbal section), the Henmon-Nelson Test of Mental Ability and the Progressive Matrices (1947). This last test was intended to provide a scale that might be regarded as more culture-free than either the Otis or the Henmon-Nelson. The Wechsler Intelligence Scale for Children was administered to 82 Indian children that were in age representative of the experimental group. The Indian children obtained Otis I.Q.'s that were significantly lower (difference of 10,0 points) and slightly less variable than those obtained by the control population of white children (176,p.34). On the Henmon-Nelson the Indian children obtained I.Q.'s that were, on the average,

lower by 20.2 points than those obtained by the control group (176,p.34). It is interesting to note that on the Progressive Matrices test no significant difference in the means between Indian and white children was obtained for any grade (176,p.37). While the Indian children obtained an average I.Q. of 96.7 on the performance scale of the Wechsler, their mean verbal I.Q.'s was only 85.6 (176, p.39). This apparent verbal deficiency was most pronounced on the vocabulary, information and comprehension subtests. In an analysis of the results on the performance scale the authors reported that the Indian children made lower than average mean scores on the picture arrangement and coding subtests; and they made higher than average scores on the picture-completion subtests (176, p.39).

It is interesting to note that studies (57,pp.341-348) (153,pp.11-15) (88,pp.50-63) with the Goodenough Draw-a-man test have shown that Indian children score better on this test than on most performance scales and their performance is superior to the published norms for white children. In certain tribes, the boys scored significantly higher than the girls and this superiority is interpreted in terms of the environment and cultural traditions (57,p.347) (88,p.62). Britton (25,pp.44-51) gave five tests of intelligence to 102 boys and 130 girls

whose social status in their midwestern community had been determined by using Warner's (180,pp.121-130) Index of Status Characteristics. He demonstrated that performance on the Goodenough scale was uninfluenced by the child's membership in a particular social class and that, at the same time, relationships between social status and I.Q. on the other four tests were moderate and statistically significant (25,pp.48-49).

Most of the comparisons of the abilities of Indians and white have been made with tests based on white culture and standardized on white groups. DuBois (61,p. 523) reversed this procedure and standardized his draw-a-horse test on Pueblo Indian children. He claims that the horse-drawing test has greater validity as a measure of mental ability for these Indian children than the Goodenough test. When both tests were given to white and Indian children, the whites excelled on the man-drawing test and the Indians on the horse-drawing test (7,p.741).

Many studies have been made with other racial groups of the effect of language handicap upon test performance. The common findings are that when children are given a verbal intelligence test those from foreign-speaking homes generally make a poorer showing as a group, and the effect of a language handicap is likely to be most serious when that handicap is present in a mild degree

(7,p.717). With the Otis group test, Mead (119,pp.465-468) found the performance of children of Italian parentage to be inferior to that of the American children. There was a consistent rise in average score with the increase in the amount of English spoken at home.

Goodenough (81,pp.388-397) administered the Goodenough Intelligence Test for Young Children (nonverbal) to 2457 public school children of various racial stocks. She found that the rank-order of the various nationality groups corresponded closely to that found by means of other intelligence tests. She also found a correlation of -.75 between the average I.Q. of children in the various immigrant groups and the tendency of such groups to retain their own language in the home. A correlation coefficient of this order would suggest that children in those immigrant groups that do not adopt the English language readily tend to obtain lower scores on our intelligence tests. Two explanations are possible; either the lower intelligence tests score is the result of the greater language handicap, or that for those immigrant groups that are slow to learn English their slowness is a direct result of their lower intellectual potentiality and poorer adaptability (7,p.718).

Berry (21,pp.185-203) gave the Detroit Primary intelligence test to 10,000 first-graders and found that

children from non-English-speaking homes tested lower than those from English speaking homes, and, of the former group, Germans tested the highest and Italians the lowest (21,p.203).

Several studies have shown that the inferiority of the immigrant groups is greatly diminished and may disappear entirely when nonlanguage tests are used. Pintner (131,pp.292-295) found that while only 37 per cent of some 165 children of foreign parents reached the median score for children of American-born parents on the National Intelligence Test (which is predominantly verbal), on the other hand, the two groups of children had identical median scores and very similar curves of distribution on the Pintner Non-Language Scale. Pintner has suggested that caution must be exercised in drawing conclusions as to the intelligence of foreign children when tested solely with verbal tests (131,p.295).

Darcy (50,pp.21-44) has shown with 212 preschool Children that while monoglots excell over bilingual subjects on the Stanford-Binet, they are inferior to them on the Atkins Object-fitting Test. She has concluded that the bilingual subjects had suffered a language handicap in their performance on the Stanford-Binet scale (50,p.41).

More recently, Darcy (51,pp.499-506) has administered the Pintner tests to 235 bilingual children of

Puerto Rican parentage. On the Pintner verbal test the mean I.Q. for this group was 79.56 while on the nonverbal test the mean was 87.84. The Pearson coefficient of correlation between I.Q.'s on the Pintner verbal test and the Pintner nonlanguage test was .58±.03 (51,p.504). Darcy has concluded that the two tests were measuring the same functions to a fairly large extent but not to so great an extent as to warrant the substitution of one test for the other. She has suggested that a combination of both types of tests would yield a more valid picture of the intelligence of a bilingual population than either a verbal or nonlanguage test administered as a sole means of intelligence measurement (51,p.506).

In a careful statistical study of the effects of bilingualism, Anastasi and Cordova (6,pp.1-19) administered the Cattell Culture Free Intelligence Test to 176 Puerto Rican children in grades six to eight of a parochial school in New York. One half of the group received the test instructions in English during the first test (Form A) and in Spanish for the second (Form B); the order of the languages being reversed for the other half of the group. The split-half reliability of Forms A and B in the English and Spanish versions ranged from .84 to .92 (6, p.9). Speed played a negligible part in the scores obtained. An analysis of variance was conducted on 108 of

the subjects including twenty-seven boys and twenty-seven girls in each of the two subgroups. F-ratios were significant for subjects, session, and interaction of order X sex. The most conspicuous finding was the marked improvement from first to second session, regardless of the language. Girls performed better when the testing order was Spanish-English, while the boys responded more favorably to the English-Spanish order. The authors have suggested that this difference may be due to the fact that the girls are less Americanized than the boys and achieve better rapport with an examiner who initially speaks their native language. As a group these Puerto Rican children fell well below the test norms. The authors have explained this discrepancy in terms of the very low socioeconomic level of the Puerto Ricans, their bilingualism, and their deficiency in both languages, together with their lack of test sophistication, and their poor emotional adjustment to the school situation. (6,p.17)

Darsie (52,pp.1-89) made an extensive study of the mental capacity of Japanese children. He gave the Stanford Binet and the Army Beta tests to 570 American-born Japanese children between the ages of ten and fifteen years and for whom English was the most familiar language. While on the Stanford-Binet test the median I.Q. of the Japanese children was 89.5, just ten points less than that for white

children of the same districts, there was no consistent difference in score on the Army Beta test between Japanese and American children. The author demonstrated that the inferiority of the Japanese with the Stanford-Binet was limited mainly to its linguistic elements (52,p.84).

Studies (157,pp.445-449) (45,pp.14-15) (35,pp. 544-551) have demonstrated that Mexicans are unduly penalized when their intelligence is rated solely by a verbal test that has been standardized on American whites. Shotwell (157,pp.445-449) and Cook and Arthur (45,pp.14-15) found that the Arthur Point Scale of Performance Tests revealed a high degree of potential ability that was not indicated by the Stanford-Binet test.

It is interesting to note that when a child speaks one language at home and another at school, his vocabulary in each language is restricted. In an extensive investigation of the intelligence of Welsh children, Saer (154, pp.25-38) found that the range of vocabulary of monoglot children was greater than that of bilinguists either in Welsh or English. He found that monoglot children showed a considerable superiority, too, on the Binet scale of intelligence (154,p.38). Barke and Williams (18,p.249) (19,pp.63-77) found the bilingual children in Wales to be inferior to monoglots on three verbal tests but the difference between the groups on a nonlanguage scale was

insignificant. Further, when Welsh forms of the verbal tests were administered the inferiority of the bilinguals was even greater than it had been on the English forms (19,p.76).

More recently, Jones (95,pp.114-123) found the mean verbal I.Q. of Welsh speaking children to be significantly lower than their mean nonverbal I.Q. owing to their inadequate reading ability in English. It also appeared that the difference observed between the two means tended to diminish as the reading age in English increased, although the gap was not entirely closed even with a reading age as high as 11 years (95,p.121).

In subsequent research, the same author (96,pp. 114-120) demonstrated that even after full allowance has been made for inferior reading ability, a group verbal test in English may not give an accurate assessment of the intelligence of Welsh speaking children and that the Welsh children are not handicapped in verbal reasoning that can be carried on through the medium of Welsh.

The writer is inclined to agree with Darcy's generalization that

> "In the more carefully controlled investigations into the effect of bilingualism upon the measurement of the intelligence of children, bilingualists achieve scores significantly inferior to those achieved by matched monoglots on verbal tests of intelligence, whereas on nonverbal tests such

> inferiority of the bilingualists has not been indicated." (51,p.499).

There are some noteworthy studies that arrived at a somewhat different conclusion. Arsenian (10,pp.1-164) conducted an extensive and carefully controlled investigation in Brooklyn of the relationship between bilingualism and mental development of 2778 children from one predominantly Jewish neighborhood and another predominantly Italian district. The Hoffman Bilingual Schedule was used to measure the extent of bilingual background, the Sims Score Card to determine socioeconomic status, the Pintner Non-Language Test and the Spearman Visual Perception Test to evaluate intellectual ability, and a comparison of the modal age-grade status in New York elementary schools was made to ascertain the age-grade status. Arsenian found no significant relationship between bilingual background and intelligence as measured by these tests. Similarly, the same author working with Pintner (133,pp.255-263) found a correlation of $-0.059 \pm .031$ between bilingualism (as measured by the Hoffman Bilingual Schedule) and scores on a verbal test for a group of 469 native born Jewish children in Brooklyn (133,p.258). The authors concluded that bilingualism per se did not influence, favorably or unfavorably, the mental development of bilingual children studied in this investigation.

A similar finding resulted from a sociological study (167,pp.371-375) in Ceylon where 212 university entrants were given the California Test of Mental Maturity (Advanced, short form). The median centile scores for the male and female students were 76.8 and 71.5, respectively, on the verbal subtests and 12.0 and 5.6, respectively, on the nonverbal subtests (167,p.372). Both men and women were above the American average on language factors, but much below it on nonlanguage factors. The author concluded that bilingualism does not necessarily exert a depressing effect on the acquisition of those skills sampled by standard intelligence tests (167,p.374). His interpretation of these results postulates the existence of an integrated culture-complex which defines the role behaviors that characterize the stratum in Ceylonese culture from which university entrants come.

A similar cultural interpretation is made by Klineberg (100,pp.173-174) to explain the superiority of Jewish subjects on verbal items. He cites a study made by Halpern in which the mean I.Q. earned by a group of Jewish children on the Stanford-Binet was 96.2 while on the Pintner-Paterson scale their average was only 81.5, and adds,

> "There is among Jewish families such a marked emphasis on schooling and upon abstract intelligence to the almost total disregard of manual

> dexterity and mechanical intelligence that this result was really to be expected." (100,p.174)

Similarly, other studies (20,pp.1-105) (143,p.221) have confirmed that Jewish children excell on verbal tests.

The instability of I.Q.'s for children of foreign-born parents has been demonstrated by Goldin and Rothschild (79,pp.673-676) who tested Italian children in grades one, four, six and eight. The coefficient of reliability between grades one and four was only 0.46 while between grades 4 and 6, 4 and 8, and 6 and 8, the coefficients exceeded 0.80 (79,p.675). They attributed the instability of mental measurements made in the early grades to language handicaps and environmental conditions (79, p.676).

It has been demonstrated that when a nonverbal test is used, the choice of language for giving directions may affect the scores. Mitchell (121,pp.29-37) administered the Otis Primary Group Intelligence Test to 236 Spanish speaking children. The mean I.Q. when instructions were given in Spanish was approximately 10 points higher than when these instructions were given in English (121,p.33).

It is generally recognized that intelligence test performance is greatly affected by differences in schooling. A survey was conducted by Garth, Lovelady and Smith (73,pp.431-435) to determine what weight should be

assigned to educational achievement in studying the intelligence of Southern Negro children. The subjects were 2006 Negro children in urban schools of Texas and Oklahoma. A combined achievement and intelligence test (Otis Classification) was administered. When intelligence was correlated with school grade (achievement held constant) r = .781. When intelligence was correlated with achievement (grade held constant) r = .781. When intelligence was correlated with both of these factors of education combined, r = .812. The authors point out that there is little left for any factor other than those of schooling (73,p.435).

Thus, there have been numerous studies that have shown the existence of group differences in measured intelligence. The earliest investigations were misguided attempts at racial comparisons and these were criticized (155,pp.765-772) for the misapplication of verbal tests. More recent studies have indicated that intelligence test scores made by immigrant children, particularly on verbal group tests, are subject to vitiation by the factors of language, schooling, social status, cultural background of the subject, and the experiential bias of the test (7,pp. 713-742). The validity of the traditional tests with foreign subjects has been seriously questioned. No longer are they considered to provide a just evaluation of the

innate capacity of an immigrant subject. For this reason, attempts have been made to design suitable instruments for this purpose. The research studies cited suggest that while cultural factors have not been completely removed from intelligence testing, they have been reduced substantially.

CHAPTER III

PROCEDURES USED IN THIS INVESTIGATION

Selection of the Experimental Group

For this study, the experimental group consisted of those German children ten to fifteen years of age who were newly enrolled in the "English for New Canadians" classes of Vancouver schools during the fall and winter terms 1954 - 1955. In all cases these immigrants had arrived in Canada subsequent to June, 1954. Without exception, their native language was German and they were extremely deficient in English. While none of the immigrants had been in Canada longer than six months at the time of testing, for most of them, their period of residence in Canada was considerably less.

At the outset of the testing project on October 27, 1954, about one-half of the experimental group were in "New Canadian" classes. For sake of convenience, these were tested in an arbitrary order of class groups. The remainder were tested in the order of their application for admission to these special classes. These

applications were made soon after their arrival in Vancouver because of the close liaison between the Department of Immigration and the Vancouver School Board. Testing of "New Canadians" ceased when there had come to be fifty subjects, (23 boys and 27 girls), in the experimental group.

It should be remembered that teachers of "English for New Canadians" seek to place these students in the regular school classes as soon as they feel that the pupils are ready. Of the individuals in the experimental group three were given promotions before testing had been completed.

While other pupils were tested at the request of the class teacher, only the data for "New Canadians" as defined on page 12, were included in this study.

Selection of the Control Group

The Progressive Matrices test was administered as a group test to more than 500 selected pupils, ten to fifteen years of age, in regular classes of Vancouver schools. From these were chosen fifty English speaking second-generation Canadians each of whom matched a German immigrant on the bases of sex, chronological age and score on the Progressive Matrices test. (See tables IV and V, Appendix B)- For forty-nine of the fifty matched pairs,

the difference in the chronological ages was four months or less and the difference in test scores was fewer than four points. The mean difference in age was 1.56 months and the mean difference in score was 1.06 points. The mean algebraic difference in chronological age was less than seven-tenths of a month and the corresponding mean algebraic difference in Progressive Matrices scores between the matched pairs was only +.06 points; both in favor of the control group.

Selection of Instruments

The design of this experiment required the following instruments:

(1) a test that could be used as a reasonably equitable basis for matching Canadians with Germans,

(2) a verbal German culture test for use with the immigrant children to provide a base from which comparisons of their performance on other tests might be made,

(3) a verbal American culture test for use with the Canadian subjects and to which their performance on culture-free instruments might be related,

(4) the performance subtests of an individual intelligence scale based on the American culture,

(5) an individual culture-free test, and

(6) a group culture-free test.

The writer has made a careful and extensive survey

of tests available in each of these categories. The following criteria were the basis of selection:

(1) each test had to be accompanied by reasonably adequate validation data,

(2) evidence had to be presented that it was a reliable instrument,

(3) the test had to be attractive and interesting, yet at the same time objective and fair,

(4) it had to have the power to discriminate adequately within the range of abilities of both groups,

(5) each test had to be sufficiently comprehensive to be an adequate test of problem-solving ability in varied situations,

(6) it had to be long enough to give consistent results, yet not too time-consuming,

(7) the test had to be easy to administer and score

(8) instructions had to be simple and few in number and, for some of the tests, the directions had to be capable of being presented in pantomime,

(9) the level of difficulty of the test items, notably the vocabulary level of verbal test items, had to be suited to the range of abilities of children ten to fifteen years of age,

(10) the norms for the test had to be appropriate to the groups being tested and the size of the standard-

ization samples had to be large,

(11) the test must have proven itself to be satisfactory in practice and, preferably, to have been generally accepted and widely used, and

(12) the test had to be available for the use of the writer in Canada at the time of this research and at a reasonable cost.

While no one test satisfied all of these criteria perfectly, each test selected appeared to the writer more nearly to meet the requirements than any other test of its type.

Progressive Matrices (1938) Test

Raven's Progressive Matrices (1938) was chosen as the basis for matching Canadians with Germans. This instrument is admirably suited to its purpose,-- it is entirely independent of language, it can be administered as a group test, it is untimed, it is used internationally, it is well standardized, it is a highly convenient tool, being simple to administer and score, and it attracts and holds the interest of most subjects. Each item consists of a design or matrix from which a part has been removed. The subject is required to examine the matrix and to select from several pieces given below it, the right one to complete it. Matrix tests have been successfully

administered to physically and mentally defective children (148,pp.40-43). The scale can be given as an individual, a self-administered, or as a group test and it is suitable for use with children above five years of age and adults. The test consists of five sets of twelve problems each, progressively graded in difficulty both between and within sets and of sufficient range of complexity to discriminate effectively among subjects in a sample of the general population. The sequence of the problems in the series provides the standard training which each person receives. A subject's total score is taken to be an index of his intellectual capacity that is independent of his nationality, language ability or education. The series provides a five point percentile grading irrespective of the subject's age. The contribution which each of the five sets makes to the total provides a means of assessing the consistency of the estimate. The author claims that this score pattern has considerable psychological significance (147,p.1) but data to support this claim are lacking.

Begabungstest B-1

The search for a well-standardized, verbal group test based on German culture and in the German language resulted in the selection of the Begabungstest B-1 (Form A). It was developed in the late 1920's by Hylla and Bobertag as a test of general intellectual ability laying

special stress on verbal elements. The test was widely used and on the basis of practical experience the test was revised several times between 1949 and 1952 by Hylla, Kunze, Durost, Ledig, Winter and Preston (140,pp.381-386). During this time, their prototype test of 200 items was administered to some 20,000 students from ten to sixteen years of age. On the basis of these results, two forms of equal difficulty were standardized for children of these ages and 15,000 tests of each form were administered. From these, 5000 of each form were selected at random for item analysis. This resulted in 28 items being discarded and the remainder were presented in two parallel forms equivalent in difficulty and identical in format, the items in each being arranged in order of increasing difficulty.

The test items are similar in content and structure to some of the better-known American verbal tests such as the Otis. Included in the test are analogies, combinations, abstractions, sequences, proverbs, information, block counting, logical inference and vocabulary items. Each question is in the form of a multiple choice item with five alternatives, the correct response to be underlined. No correction is made for guessing. The time for the test is forty minutes. There are provided standardization data based on the results of the test with 4,000 boys and girls from ten to fifteen years of age

selected from four representative parts of Berlin. Separate norms are provided for boys and girls and these are expressed in standard scores rather than in percentiles. For each age range the mean is one hundred and the standard deviation is fifteen points.

The manual presents evidence of the test's validity. Teachers were asked to estimate the intellectual capacity of their students. Test scores were correlated with these estimates; for grade six pupils r = +.66; for pupils in grades seven to nine, the coefficients varied from +.30 for those following the "practical" course to +.45 for those following the "academic" program and to +.47 for those receiving "technical" education. The odd-even reliability coefficients ranged from +.76 to +.85.

In the course of selecting a German test the writer exchanged communications with Dr. A. Schwarzlose, Berlin Lichterfelde and with Dr. H. M. Ledig, of the Schulpsychologische Arbeitstelle, Berlin, where this test was developed. Both of these correspondents have recommended the use of Begabungstest B-1 in this experiment. In a letter from Berlin - Lankwitz (dated October 22, 1954) Ledig reported that good correlations were found between the Cattell test (Forms 2A and 2B) and the Begabungstest B-1 when they were administered to 1000 students in grades five to eight of the schools in Berlin. Dr. Ledig

expresses the opinion that cultural background would certainly not affect the score on the Cattell test so much as it would on a verbal test, but at the same time he feels that the effects of scholastic training are not eliminated from "culture-free" instruments.

Ledig points out that the Begabungstest is known to the students of Germany by the name "Wer kann gut nachdenken?" and it is by this name that Preston (140,p.383) and Froelich (69,pp.568-573) refer to the test in their summaries of psychological testing in West Germany.

Consideration was given to Amthauer's "Intelligence-Structure Test" (5,pp.1-38). It has two notable features, occupational norms and provision for deriving equivalent Wechsler I.Q.'s. The writer chose not to use the Intelligence-Structure-Test because it appeared to be suited more to adults than to children. No norms were available below the age of fourteen years.

Preston (140,pp.383-384) has described the "Testheft B" developed for use with sixth grade pupils of Hamburg in 1950. He reported that the test is too difficult for pupils of the sixth grade and that its standardization is incomplete.

Otis Self-Administering Test of Mental Ability

The Otis Self-administering Test of Mental Ability: Intermediate Examination: Form A was selected as

the verbal American culture-test to be administered to the Canadian group in this experiment because of all such tests it is most like the Begabungstest which was given to the German children. The format is almost identical. Most of the seventy-five items are of the multiple-choice type each with five alternatives and the correct one is to be underlined. No correction is made for guessing. There is a striking similarity of content in the two tests. The Otis includes many forms of vocabulary items, analogies, classifications, information, number series, opposites, proverbs, logical inferences, and arithmetic problems. In his manual (127,p.12) Otis reports validity coefficients ranging from .55 to .59 between test scores and scholarship and an average coefficient of correlation of .842 between the Intermediate and the Higher examinations. Stalnaker (164,p.61) reported correlations ranging from .55 to .62 between Otis scores and achievement in school subjects. Otis (127,p.12) reported an average reliability coefficient between Forms A and B of +.948. Standardization data were provided on the basis of scores made by 60,000 pupils in grades six and eight. From the raw score, the Binet mental age equivalents, centile rank, and Otis I.Q. may be read from tables of normative data provided.

The self-administering tests were modeled after a group test of mental ability designed in 1918 for use in

a large commercial establishment in Connecticut (127,p1). The Otis tests have been revised, improved and widely used with the passing years. Because of their easy administration and simplified scoring they have continued to be popular verbal tests for use in Canadian schools.

Wechsler-Bellevue Intelligence Scale

The Wechsler-Bellevue was selected as the performance test based on American culture that most nearly suited the needs of this research. The full scale is an individually administered point scale that is particularly suitable for appraising selected verbal and nonverbal mental abilities of individuals from ten to seventy years of age. Form I was published in 1939 and in the intervening years it has been generally accepted in clinical practice, particularly in the United States of America. Related research with the Wechsler-Bellevue intelligence scale has been reviewed on pages 42-45 of the previous chapter. The scale contains eleven subtests, six verbal, and five performance tests. These tests may be combined to form four separate but interrelated intelligence scales, as follows: an individual adult examination for ages 16 to 60; an adolescent scale for ages 10 to 16, consisting of the same tests but separately standardized; a performance scale consisting of five tests; and a verbal scale consisting of five or six tests depending on whether or not the

vocabulary test is included. The adolescent scale was chosen rather than the adult since its standardization is more appropriate to the age-range of the subjects included in this research project.

For the purposes of this project, only the non-verbal subtests have been used since the subjects' lack of facility with English would invalidate the results of any verbal test administered to new Canadians. The five performance tests require the manipulation of concrete materials.

The Picture Arrangement test consists of a series of pictures which, when placed in the right sequence, tell a story. The pictures are presented in a disarranged order and the subject is asked to put them in the right order so that they tell a sensible story. The situations depicted are human and practical and they appeal to most subjects. Probably the most favorable feature of the Picture Arrangement test is that it requires the subject to comprehend the total situation before he can attempt a solution. Administration and scoring are relatively easy.

The coefficient of correlation between the score on the Picture Arrangement test and the total score is quoted to be +0.51 (182,p.89).

The second test is the Picture Completion Test which requires the individual to discover the missing part

in each of fifteen incomplete pictures. He must differentiate essential from unessential details. Apparently, it is a measure of the subject's basic perceptual and conceptual abilities. The Picture Completion is a popular test and it is extremely simple to administer. The coefficient of correlation of the score on this test with the total score is quoted to be +0.61 (182,p.91).

The Block Design test is a modification of Kohs' original test who offered it as a comprehensive measure of nonverbal intelligence. In this test the subject is required to make seven designs with colored blocks that are the same as those on seven design cards. The performance is scored for both accuracy and speed. It would seem to the writer that the Block Design test is a measure of the subjects' ability to perceive a design, to analyse the pattern, and to synthesize its components, and that it is weighted considerably by his ability to solve problems in spatial relations.

The Block Design test is the best single performance test in the scale and it correlates highly ($r = .73$ for ages 35 - 49) with total score (182,p.94). Wechsler claims that it is one of the few performance tests that seemingly does measure very much the same sort of thing that verbal tests measure, and that it correlates better with Comprehension, Information and Vocabulary than some

of the verbal tests themselves. (182,p.92)

The Digit Symbol test is a form of the substitution or association test found in many intelligence scales. The subject is required to associate a certain symbol with each of ten Arabic digits and to enter the appropriate symbol in the space below each digit. There are sixty-seven digits arranged in scrambled sequence and the subject is required to substitute symbols for as many of these as he can within the time limit of one and one-half minutes. The question of what the digit symbol test measures has long been a subject of disagreement. Burik (29,pp.33-42) has shown that it involves both associational learning ability and hand-eye coordination and that of these two, the motor factor predominates. Wechsler has interpreted the speed and accuracy of the subject's performance to be a measure of his mental ability. (182,p.94) The Digit Symbol test is easily administered and it appears to be enjoyed by adolescents. The test correlates well with total score. (182,p.96)

The Object Assembly test consists of three separate items; the manikin, the feature profile, and the hand, presented in that order. The pieces of each object are arranged before the subject and he is required to put them together correctly as quickly as he can. The manikin test is scored for accuracy alone, but for the other two items

credit is also given for speed.

The Object Assembly test is intended to measure perceptive ability and insight into spatial relationships of familiar objects. It correlates poorly with most of the other subtests and with total score (182,p.98).

Each of the raw scores on the five tests is converted to an equivalent weighted score. From a table of normative data appropriate to the age of the individual, the total weighted score is converted to a performance I.Q.

Arthur Point Scale of Performance Tests

The individual culture-free test selected was the Arthur Point Scale of Performance Tests (Revised Form II). This test is designed to be used as a nonverbal measure of intellectual capacity. While its chief value lies in supplementing verbal intelligence test ratings, the Arthur scale is especially useful in cases where verbal tests are inadequate because of speech or hearing defects, reading disabilities, cultural differences or language handicaps.

The scale is composed of five subtests.

(1) In the revised Knox Cube test, four one-inch cubes are fastened on a base in a row two inches apart. The examiner taps the four cubes in a certain order and the subject is required to copy the performance. Eighteen different tapping patterns are given. The number of series

repeated correctly constitutes the score. A second trial of the Knox cube test is given after the Porteus Maze test and the raw score for the test is the average of the two trials.

(2) The Seguin form board test consists of a board containing ten variously shaped recesses (circle, square, rectangle, semi-circle, cross, triangle, star, diamond, oval and hexagon), into which correspondingly shaped blocks are to be fitted. The subject is given three trials in which he must place the blocks in the correct holes as quickly as possible. The score is the time in seconds for the fastest of the three trials.

(3) The Arthur stencil design test requires the subject to reproduce in form and color a design in two to six colors from the supply of six square cards and twelve stencils provided. There are twenty increasingly complex designs to be reproduced and the score is the number of designs that are completed correctly each within a four-minute time limit.

(4) The Porteus maze test (Arthur revision) consists of fourteen different mazes arranged in order of increasing difficulty. The subject is required to start at the beginning of each maze and find his way out. He marks his progress with a pencil and no retracings are allowed. Two or four trials are permitted depending on the maze.

The score is the number of points earned according to Arthur's schedule of credits for each maze.

(5) The Healy picture completion test. Arthur (16,p,21) suggests that this test is inappropriate for use with individuals whose cultural background is different from that of the average American urban school child. For this reason, the writer chose to exclude the Healy picture completion test from this investigation.

For each of the four subtests, the raw score has a corresponding point value and the total point score on the four tests is converted to mental age from normative data based on the scores of 968 pupils 4.5 to 15.5 years of age, from an American middle-class district. The mental age is then divided by chronological age to provide Arthur I.Q.'s.

IPAT Test of "g": Culture-free

The IPAT, test of "g" was chosen to be the culture-free group test employed in this study. This is a perceptual instrument requiring visual discrimination and the eduction of relations among the nonverbal geometrical symbols. It is available at three different levels. Scale 1, designed for children four to eight years of age and defective institutionalized adults, involves pictorial materials and rather specific verbal directions. The

writer believes that for this reason, this scale is not entirely "culture-free". Scale 2, which is used in this study, was intended for use with eight to thirteen-year-olds and unselected adults. Scale 3 was designed for upper high school and college students and superior adults. These two scales are very similar to the original Cattell test except that the subtest Pool Reflections has been replaced by a subtest called Conditions which is a more complex relation-finding test designed to reduce the weight of spatial ability in the total score. The 46 items of scale 2 are grouped into four subtests according to the relationship existing among the elements of the problem; series, classifications, matrices and conditions. A generous practice session is given at the beginning of each subtest to reduce the effects of differential test sophistication and the items are graded in order of increasing difficulty. The total raw score can be converted into 'classical' or standard-score I.Q.'s on the basis of normative data on scores of 713 pupils in two midwestern university towns and 2,584 pupils in an industrial city of Britain (42,p.8) (40,p.156).

It should be noted that the publication of Cattell's original culture-free test has been discontinued and hence it was not available to the writer. Furthermore, the test was intended for use with pupils in grades 9 to

16 and with adults (31,p.384), hence it is not a suitable test for subjects ten to fifteen years of age.

Other Tests Considered

Other tests were considered for use in this project. The Davis-Eells Games promises to be particularly useful for comparisons of people from different socioeconomic levels within one culture. The test is designed so that performance will be largely independent of reading skill, inschool instruction, or speed of response. While the problem situations, pictures, vocabulary, and syntax are those common to urban groups in the American culture, the test is not intended for use with persons of foreign cultures. Furthermore, neither the Primary nor the Elementary level is suited to the age range of the subjects in this experiment.

The Leiter International Performance scale that has been described on pages 33 and 34 was given serious consideration largely because it employs simple techniques and attractive materials in a unique situation. However, the writer rejected it because of its relatively high cost, the all-or-none aspect of its scoring system, the large amount of material to be handled, and Tate's findings that it is no more culture-free than the Arthur scale and that its norms are in need of revision (168,pp.497-501).

The Terman-McCall-Lorge Nonlanguage Multi-Mental test was first prepared by Terman for use in a national survey of education in China under McCall's supervision. On the basis of their experience, the authors recommend its use with people who do not speak English. However, the content consists entirely of drawings and for this reason the test was not chosen because pictorial symbols involve cultural influences.

The Chicago nonverbal examination was designed for use with children who are deaf, who come from homes where a foreign language is spoken or who are deficient in the use of the English language. As an attempt to reduce the cultural contamination the test has not been unsuccessful. The results on it are reported to have compared favorably with those obtained on the Pintner Nonlanguage Mental Test, the Revised Beta Examination and others, (30,p.212). The Chicago Nonverbal Examination was not selected because it has not always been found to be adequately discriminative for the age-range of subjects in this investigation (124,p.45) and because the drawings in it have been found to be unsatisfactory (30,p.213).

Serious consideration was given to the abstract reasoning test in the Differential Aptitude Test Battery. In content this test closely resembles the original Cattell culture-free test. It involves the detection of a

principle of change occurring in four successive geometrical figures and then the selection of an appropriate fifth figure to be chosen from a group of five alternatives. While the test has been carefully developed and standardized, it was designed for students in grades eight to twelve, and it would be too difficult for younger children included in this study.

Mode of Administration of the Tests

In the administration of the tests, the writer has endeavored to follow the standardized procedures for each test, to establish and maintain good rapport with the subject so that he will put forth his best effort, and to exercise diligence and care in the scoring of responses and in the recording of results.

The Begabungstest, Otis, Progressive Matrices, and Cattell tests were used as group tests; the Wechsler-Bellevue and the Arthur were given individually. In all cases, the test manuals were carefully studied and standardized instructions were followed without deviation. The writer gave all of the tests except the Begabungstest on a trial basis to other subjects before employing them in the present study. The writer chose to initiate the testing of German subjects with the Begabungstest with the belief that they would have a greater sense of security in their native language and that they would gain confidence by

having that test first. In a second testing session the Cattell test and Progressive Matrices were administered. The Cattell test was given first only because it is a timed test and the Progressive Matrices untimed. A simple mimeographed answer sheet for the Progressive Matrices proved to be entirely satisfactory. A convenient cardboard mask enabled the writer to score the test with facility. For the Cattell test, printed answer sheets were obtained from the test publisher but these proved in the trial run to be unsatisfactory because of the poor format and the difficulty encountered in instructing subjects as to their use. Hence, the writer reverted to using the test booklet alone. In a third session, the individual tests were given alternating the order of the Arthur and the Wechsler-Bellevue scales. The writer was careful to declare a recess between the tests and for young subjects other rest periods were taken when there was any evidence of fatigue or lagging interest. For the Arthur, the subtests were given in the following order: Knox Cube, Seguin Form Board, Stencil Design, Porteus Maze, and Knox Cube (second trial). The writer found that with the Wechsler the Object Assembly was generally satisfactory as the initial test. Then the Picture Completion, Picture Arrangement, Digit Symbol and Block Design test, in that order seemed best to facilitate administration.

Further, for some children their failure to complete satisfactorily the block designs was frustrating and for this reason the test was deliberately left till last.

For Canadian subjects the Otis was the first test, otherwise the sequence of testing was identical to that used with the immigrant children.

Group tests were administered in the classroom of the group being tested because the subjects were more likely to feel at ease there. Physical conditions were adequate. For the individual tests, the subject was alone with the examiner in almost all cases. Only on a very few occasions were there interruptions and these were minor and unavoidable.

The Wechsler-Bellevue, the Arthur, the Progressive Matrices and the Cattell tests were administered by the writer to all of the 100 subjects. A few of the German subjects wrote the Begabungstest (which is a self-administering test) under the supervision of their teacher. A number of the Canadian students in grades six and seven had recently been given the Otis test by the school psychometrist. In these cases, the results were obtained from the bureau of measurements and the test was not repeated. For all other subjects these verbal tests, too, were administered by the writer.

Before the project was undertaken, the cooperation of school officials and teachers was obtained and this contributed largely to the success of the project. In all cases the classroom teacher was quick to display her acceptance of the examiner and this was promptly reflected in the reaction of the pupils. The writer was impressed by the genuine desire of the subjects to take the tests and by the interest and enthusiasm with which they performed. For his part, the writer endeavored to establish and maintain an effective rapport with each subject. The examiner was careful to avoid a display of dissatisfaction with an inferior response. Throughout the research he has endeavored to maintain a balance between a judicious sensitivity to the reactions of the subject and a careful regard for the scientific demands of psychometric instruments.

The difficulty of administering tests to subjects with a language handicap was much less than anticipated. When it was necessary an interpreter translated the instructions for the group tests. When the time came to administer the individual tests, most of the subjects had acquired a sufficient understanding of English that the few simple instructions could be given verbally. Where there was any doubt, these were presented in pantomime. In any case, the subjects seemed to have no difficulty in

understanding the directions.

Statistical Techniques Employed

Before applying statistical techniques or making any comparisons, the test results had to be reduced to a common scale. To this end the raw scores on the Begabungstest, Otis, Cattell and Wechsler-Bellevue tests were converted directly to intelligence quotients from standardization data based on large samples. It is interesting to note that for the German verbal test separate norms are provided for boys and girls, the standard being higher for boys. It may be that this difference is a reflection of the separation by sexes in the organization of German schools and the apparently greater academic provision for male students. For the Arthur scale raw scores were converted to equivalent mental ages and from these intelligence quotients were computed.

These intelligence quotients did not provide an equitable basis for comparison because of the widely different values of the standard deviation. For this reason, all of the intelligence quotients were changed to standard scores (z-scores) with a mean of 50 and a standard deviation of 10 points. These standard scores were the bases for comparisons between groups and among tests.

The raw scores on the Progressive Matrices test were used as a basis for matching the individuals.

However, the test manual presents a table of percentile values for raw scores of subjects at various age levels. Each percentile score in Raven's table has been assigned by the writer the standard score which it represents in a normal distribution (T-scores) and for intermediate values, the writer has had to interpolate, although this represents an approximation. This is precisely the method currently being used in the Institute of Psychiatry, Maudsley Hospital, London to convert Matrices scores. Another approach is to consider that the approximate percentile value of each raw score has been converted to an equivalent score in a normal distribution having a mean of 50 and a standard deviation of ten points. While T-scores and Z-scores are not interchangeable, they do correspond closely, and the more normal the original distribution the closer is the correspondence.

The standard scores on the verbal culture test, Wechsler-Bellevue, Arthur and Cattell test (columns 2, 3, 4, 5, respectively) and the normalized score on the Progressive Matrices test (column 6) are presented in tables VI, VII, VIII, and IX, Appendix B.

Mean values of these scores were computed and first order comparisons were made of the results among the tests, between the groups and between sexes within the groups. Pearsonian coefficients of correlation were computed

between results on any two of the various tests for boys and for girls separately to avoid interaction between test and sex. The formula employed here was:

$$r = \frac{N(\Sigma XY) - \Sigma X . \Sigma Y}{\sqrt{N(\Sigma X^2) - (\Sigma X)^2}\ \sqrt{N(\Sigma Y^2) - (\Sigma Y)^2}}$$

The writer has referred to one of the tables published by Lindquist (107,p.212) which shows the correlation needed to reach the 5 per cent and 1 per cent levels of confidence for varying numbers of cases and he has noted accordingly correlations that are significant at these levels.

The most significant statistical technique employed in this research is an analysis of variance, three-way classification. In this treatment of the results, the sum of squares is partitioned into its components permitting a study of the variation in test performance due to differences in nationality, sex, and test, and the interaction of these factors. This analysis has been applied to the scores of all the subjects on the verbal (Begabungstest or Otis), Wechsler-Bellevue, Arthur, and Cattell tests, and also, to the weighted-scores of all subjects on the five subtests of the Wechsler-Bellevue Performance Scale. It is regretted that a similar analysis could not be made with the subtests of the Arthur or Cattell tests for the reason that the subtests of these instruments do not con-

tribute equally to the total score. For example, the normal range of scores for subjects ten to fifteen years of age on the Stencil Design test receives a much higher point value than do corresponding scores on the Knox Cube test. The variable weighting of the subtests with age renders such an analysis difficult, if not impossible.

In considering the variation due to tests (or subtests) a rather new technique, the Duncan multiple-range test has been used. It segregates the mean scores into homogeneous groups. A detailed explanation of this statistical technique has been included in Appendix A.

Finally, an attempt has been made to compare test results with teacher-evaluations. Teachers were asked to rate the classroom performance of their students by assigning letter grades A (top 5%), B (next 20%), C (middle 50%), D (next 20%) and E (lowest 5%). The statistical result here was not entirely satisfactory because of the preponderance of "C" ratings and the rarity of A's and E's. These letter grades are indicated in column 7 of tables VI, VII, VIII, and IX, Appendix B. An analysis of variance has been completed to test the hypothesis that there is no significant difference in scores for individuals who are assigned different ratings by teachers. A summary of the statistical results has been presented in table XXVIII, Appendix B.

CHAPTER IV

FINDINGS AND INTERPRETATIONS

The performance of all of the 100 subjects is summarized in tables VI, VII, VIII, and IX, Appendix B. Standard scores are used for the Begabungstest, Otis, Wechsler-Bellevue, Arthur and Cattell tests while for the Progressive Matrices performance is expressed in terms of equivalent scores in a normal distribution. Teachers' ratings are expressed in terms of letter grades A, B, C, D, or E. (It should be remembered that the Progressive Matrices was employed to match individuals of the control group with those of the experimental group. Consequently in analyses involving both groups, the Progressive Matrices scores cannot be included in the experimental data.)

The statistical results are presented largely in tabular form. Table X, Appendix B summarizes a three-way analysis of variance of standardized scores made by both groups on four mental tests. The variance is partitioned for variation due to nationality, for sex differences, and for variation among the four experimental tests (the culture verbal test, the Wechsler-Bellevue Performance Scale,

the Arthur Point Scale and Cattell's IPAT test of "g"). Average scores on these tests for boys and girls of both groups are presented in tables XI, XII, XIII, and XIV, Appendix B. Table XV, Appendix B shows the differences in mean scores on the four tests between the two groups for both sexes. The weighted-scores on the five performance subtests of the Wechsler-Bellevue Intelligence scale for all of the 100 subjects are presented in tables XVI, XVII, XVIII, and XIX, Appendix B. A three-way analysis of variance (table XX, Appendix B) has been made of the standardized scores on the performance subtests of the Wechsler-Bellevue Intelligence test. Here, again, the variance is partitioned for variation due to nationality, for sex differences, and for variation among the five subtests. The mean scores on the five Performance subtests of the Wechsler-Bellevue intelligence test are presented in tables XXI, XXII, XXIII, and XXIV, Appendix B. Coefficients of correlation between the tests are shown in table XXV, Appendix B. Average scores on subtests of the Arthur and Cattell tests for boys and girls of both groups appear in tables XXVI and XXVII, Appendix B, respectively. Finally, an analysis of variance is summarized in table XXVIII, Appendix B, to test the hypothesis that there is no significant difference in scores for individuals who are assigned different ratings by teachers.

The findings of this experiment are presented in terms of answers to the specific questions that were raised in chapter one.

(1) Do immigrant children (who have been matched with second generation Canadians on the bases of age, sex and raw score on the Progressive Matrices Test) score consistently lower than, higher than, or about the same as Canadians on other tests used in this research?

When we consider the variation due to nationality the 'F' value of 156.16 (table X, Appendix B.) is significant at the five per cent level, that is the German average of 50.63 points is significantly lower than the Canadian average of 56.72 points. Hence we may conclude that members of the experimental group score consistently lower than members of the control group.

(2) If a difference exists between the performance of the immigrant children and that of Canadians, for which test is this difference greatest? Or, are the nationality differences about the same for all of the tests?

From tables XII and XV, Appendix B. it will be seen that the differences between the groups were as follows: on the culture verbal test, 7.06 points; Wechsler-Bellevue Performance test, 5.80 points; Arthur Point Scale, 5.14 points; and the Cattell, 6.36 points; in all cases

the average score for Canadian group was higher than that of the German group. However, because the 'F' value of 0.70 (see table X, Appendix B.) for the interaction between tests and nationality is not significant at the five per cent level we may conclude that: (a) while the differences between Germans and Canadians in the average scores on the four tests are substantial, these nationality differences are about the same for all tests, (b) the relative average scores for the four tests are about the same for Germans and Canadians; and (c) the differences among the average scores on the four tests are about the same for each of the two national groups. These conclusions suggest that for two groups matched on the basis of their scores on the Progressive Matrices none of these tests is biased more than another in favor of one group. Since one of the tests is a culture test in the subject's native language and hence is unbiased, we may conclude that cultural factors in the other three tests, (the Cattell IPAT test, the Arthur Point Scale, and the Wechsler-Bellevue Performance Scale) have been effectively reduced.

(3) Is the average score for boys higher than, lower than or about the same as the average score for girls?

The average score for boys, 55.05, is higher than that for girls, 52.50, (table XIV, Appendix B.). In the

analysis of variance, the 'F' value 27.30 (table X, Appendix B.) for this variation due to sex is significant at the five per cent level. In a sense, this superior performance of the boys is to be expected for when the groups were established, the boys happened (by the chance selection of the experimental subjects) to exceed the girls on the Progressive Matrices test which was used for matching purposes. The boys excelled the girls most on the Wechsler (by 4.26 points) and the Arthur tests (by 3.95 points) and least on the Cattell (1.09 points) and Culture tests (0.91 points). While this apparent superiority of the male subjects may well be worthy of further investigation, it has little bearing on the fundamental purpose in the present research.

(4) Which of the two national groups has the greater variation due to sex?

(5) Is the variation in scores due to nationality greater for boys or girls?

The 'F' value of 11.92, which is significant at the five per cent level, indicates that there is interaction between sex and nationality (table X, Appendix B.). This interaction can be interpreted in different ways in this situation to answer each of the two questions:
(4) While German boys score higher than German girls, the amount of their superiority is not so great as that shown

by Canadian boys over Canadian girls (see table XIII, Appendix B.). Even though this finding may provide a basis for interesting speculation and subsequent research, it is beyond the scope of this study. (5) The variation due to nationality is greater for boys than it is for girls (see table XIII, Appendix B.). For this reason, if we are to make comparisons between nationalities, we must treat the boys' data separately from the girls' data.

(6) For boys, which tests gives a mean score not significantly different from those on other tests?

(7) For boys, which of the tests would give about the same relative ranking in a large unselected population as that given to the same group by the Progressive Matrices test?

When we consider the variation due to the tests, the 'F' value of 5.91 (table X, Appendix B.) is significant at the five per cent level. This means that the average scores on the four tests for the same sample members are not the same. However, in view of the fact that there is interaction between test and sex any conclusions about variation due to the tests must be drawn for boys and for girls separately. For the boys the mean scores were:

Wechsler	Cattell	Arthur	Culture test
56.50	55.65	54.54	53.52

With these data, we can use a five per cent level Duncan multiple range test that segregates the means into homogeneous groups. (Since this is a new test, a complete explanation of the procedure is made in Appendix A.) By this test, we find that the means underscored by the same line are not significantly different and any two means not underscored by the same line are significantly different.

For boys, only the Grace Arthur test gives a mean score that is not significantly different from those on the other tests. Furthermore, the average score (54.54) on the Arthur Performance Scale is in close agreement with that (54.78) made by boys on the Progressive Matrices Test used in setting up the national groups. This comparison is made to show that the boys group would be given about the same relative ranking in a large unselected population by both tests.

(8) Which test, if any, would give a mean score for girls that is not significantly different from the mean score on other tests?

(9) For girls, which of the tests would give about the same relative ranking in a large unselected population as that given to the same group by the Progressive Matrices test?

The mean scores for the girls were:

Cattell	Culture test	Wechsler	Arthur
54.56	52.61	52.24	50.59

For girls, the Wechsler Performance Scale gave a mean score not significantly different from that of the culture or Arthur tests. Furthermore, the average score (52.24) on the Wechsler agrees well with the average score (52.00) on the Progressive Matrices, that is, the Wechsler would give these girls about the same relative ranking in a large unselected population as the Progressive Matrices test. Only the Cattell test gave an average score significantly different from the averages on the other tests.

(10) Which, if any, of the culture-free tests gives a mean score that is significantly different from the mean score of the culture test?

Only the Cattell test gave a mean score that was significantly different from (and higher than) the mean score on the culture verbal test. While this superior performance on the Cattell applied to male and female groups alike it should be noted that for boys, the mean score on the Wechsler-Bellevue scale was also significantly higher than the average score on a culture-test, and for girls, the mean score of the Arthur was significantly below their mean on the culture-test.

(11) For which test is the difference the greatest between the mean score for Canadians and the mean

score for Germans?

The superiority of Canadians over Germans in average scores on each of the tests is summarized in table XV. However, since the analysis of variance (Table X) has revealed that the interaction between test and nationality, and the interaction between test, nationality, and sex are not significant, we must conclude these nationality differences, for boys and girls alike, are about the same for all of the tests.

(12) If there are components that operate in culture tests to make the relative standings of the national groups different from their relative standings on the Progressive Matrices test, do these test-components operate to a greater or to a lesser extent in the Cattell, Wechsler, and Arthur tests?

Because the nationality differences on the four tests are about the same, we may conclude that those test-components in the culture-tests that tend to make the relative standings of the national groups different from their relative standings on the Progressive Matrices test, must operate to approximately the same extent on all of the tests.

The fact that although the groups were matched on the Progressive Matrices test, the Germans scored below the Canadians on other tests suggests that if the groups had

been matched on the basis of their scores on culture-tests, the Germans would have excelled the Canadians on the Progressive Matrices test and would have compared more favorably on the Cattell, Wechsler and Arthur tests. Such a problem merits further investigation.

(13) What are the Pearsonian coefficients of correlation between results on any two of the tests for boys and for girls of both national groups, and are these values significant?

In view of the indicated interaction between tests and sex and the absence of interaction between tests and nationality (see table X, Appendix B) the data have been segregated so that the correlation of results could be determined separately for male and female subjects from both national groups combined. These coefficients range from .50 to .89 (see table XXV, Appendix B). Lindquist (107,p.212) has evaluated the degree of correlation needed for significance and his results indicate that all of these coefficients are significant at the one per cent level.

(14) Between which two tests is there the greatest agreement?

If we think of agreement in terms of the tendency for two measures to vary concomitantly, then the greatest agreement of results for boys is between the Wechsler and

Arthur tests (r =.89). It is interesting to note that both of these are administered individually. The poorest correlation (.50) was between the results on the Cattell and Progressive Matrices tests.

For girls, the Wechsler Performance Scale correlated best and equally well with the Arthur and the culture tests.

Generally speaking, the Wechsler appeared to correlate best with the other tests; the Cattell and Progressive Matrices, least.

If, however, we consider the extent to which two tests gave the same relative rankings in a large unselected population, then, as has been indicated in the answers to questions 6 and 7, the greatest agreement for the boys was between the Wechsler and Cattell tests; and for the girls the least difference of means was between the Wechsler and culture tests. For boys, the mean score of the Arthur was not significantly different from any of the other mean scores. For girls, the Wechsler showed agreement with all but the Cattell test.

(15) For Canadians, is the mean score on subtests of the Wechsler-Bellevue Intelligence Test higher than, lower than, or about the same as the mean score for Germans?

The scores made by the subjects on the performance

subtests of the Wechsler scale are presented in tables XVI, XVII, XVIII, and XIX, Appendix B, and a three-way analysis of variance of these scores is summarized in table XX, Appendix B. When we consider the difference in mean scores due to nationality, the 'F' value of 50.76 is significant at the five per cent level. The Canadians, with an average on all of the subtests of 9.77 points (table XXIII, Appendix B), were significantly superior to the Germans who averaged 8.48 points.

(16) How does the mean score of boys on the subtests of the Wechsler-Bellevue Intelligence Test compare with the mean score for girls?

The variation due to sex is significant at the five per cent level ('F' = 12.07); the boys scored higher than the girls (table XXII, Appendix B). This result is probably brought about by the nature of the sample. It should be noted that there is no interaction between nationality and sex; the 'F' value of 0.47 is not significant at the five per cent level. The superiority of Canadians is consistent and of similar magnitude for both sexes, and the superiority of boys over girls is consistent and similar in magnitude for both nationalities (table XXIV, Appendix B).

In considering the variation among results of the different subtests the 'F' value of 20.49 is significant

at the five per cent level. However, in view of the interaction between subtests and nationality (F= 3.12) and also between subtests and sex (F= 4.48), any conclusions about the relative performance on the various tests will have to be specific in terms of nationality or sex. (Since there is no interaction between test-nationality-sex (F= 0.72), the conclusions about test results need not be specific about both nationality and sex simultaneously.)

For the boys the mean scores were:

Object Assembly	Block Design	Picture Arrangement	Picture Completion	Digit Symbols
11.28	9.87	9.35	9.15	7.67

For boys the shortest significant ranges of 2, 3, 4, and 5 means are .82, .87, .90 and .92, respectively. Thus the mean score on the object assembly test was different from and higher than the mean scores on the other tests, while the mean score on the Digit Symbols test was different from and lower than the other mean scores.

For the girls the mean scores were:

Object Assembly	Picture Arrangement	Block Design	Digit Symbols	Picture Completion
9.96	8.85	8.57	8.44	8.35

For girls the shortest significant ranges of 2, 3, 4, and 5 means are .76, .80, .83, and .85, respectively. Again,

the mean score on the Object Assembly test was significantly higher than the mean scores on the other tests among which the differences were not significant.

For Germans the mean scores were:

Object Assembly	Block Design	Picture Arrangement	Picture Completion	Digit Symbols
10.24	8.92	7.96	7.94	7.36

For Canadians the mean scores were:

Object Assembly	Picture Arrangement	Picture Completion	Block Design	Digit Symbols
10.90	10.20	9.50	9.42	8.82

For the Germans and Canadians alike, the shortest significant ranges of 2, 3, 4 and 5 means are .79, .83, .86 and .88, respectively. Any two means not underscored by the same line are significantly different and any two means underscored by the same line are not significantly different.

(17) On which subtest of the Wechsler-Bellevue Intelligence Test did the groups perform best?

(18) On which subtest of the Wechsler-Bellevue Intelligence Test did the groups tend to perform poorly?

All groups obtained the highest scores on the object assembly test. There was a tendency for all groups to perform relatively poorly on the digit symbol and

picture completion tests.

(19) For which of these subtests was the superiority in the performance of Canadians over Germans most marked?

(20) For which of these subtests was the superiority in the performance of Canadians over Germans least marked?

The superiority in performance of Canadians over Germans was most marked on the picture arrangement and picture completion tests and least on the block design and object assembly tests.

As culture-free instruments, pictorial tests have limitations. Their validity is dependent upon the appropriateness of the pictures to the experience of the subjects being tested. In his manual, Wechsler has stated that his aim has been to choose situations from the American scene and he has admitted that, in so doing, some of the sequences may be puzzling to subjects of foreign origin (182,p.90). In this investigation, there were evidences that the New Canadians were handicapped somewhat by the cultural bias of certain items. Some of the immigrant children who had come from continental Germany had never seen a crab and they could hardly be expected to recognize that a leg was missing. The item presented little difficulty to Canadian children living on the

Pacific coast. In the picture of the adult male, many of the New Canadians failed to recognize the need for a tie to complete his attire; apparently this sartorial accessory was beyond their experience. On the other hand, for urban Canadians men's neckwear is commonplace. This difference in experience between the two groups is reflected in their responses to this item. The writer found that immigrant children did not succeed quite so often as the more sophisticated native subjects on the more difficult items, notably, the taxi and flirt sequences. The elevator series has a weakness in that the subject who cannot read English is deprived of one important cue, viz. the sign "Ringing Bell Means Rising Elevator". Because of this, several of the experimental group (and none of the control group) made the error of assuming that the elevator was descending and arranged the pictures in the reverse order.

The picture series is similar to the short comic strips to be found in the daily papers. Some children are very familiar with comics, others have had relatively little experience with them. The writer feels that the variable amount of exposure to cartoons, comics, playing cards, and picture puzzles will affect results on this test. It is interesting to note also that a larger proportion of girls than of boys correctly identified the missing

eyebrow and more boys than girls detected the missing thread at the base of the electric bulb.

In administering the tests, the writer formed the opinion that the block design test rewarded manual dexterity and that a subject's performance might be affected by the extent of his childhood experience with blocks. The digit symbol test seemed to reward to some extent hand-eye coordination and motor speed. The examiner observed that some left-handed persons might be handicapped by the present format of the digit symbol test. The object assembly test appeared to the writer to be a test of perception, and previous experience with jig-saw puzzles is likely to benefit the subject. Certainly, persistent effort is rewarded. The writer observed that on the "hand" item, particularly, subjects who became discouraged were penalized. The familiarity of the subject with the completed configuration would affect his performance. For this reason, the choice of human configurations appears wise when the subjects come from different cultural backgrounds. Another evidence of cultural bias appeared in the digit symbol test. The Germans were handicapped somewhat by digits, (notably 1 and 7) that were different from the printed style of digits in German print (1, 7). It would seem to the writer then, that tests requiring a verbal or written response fail to remove cultural bias.

Of the Wechsler scale, only the Block Design and the Object Assembly tests are entirely independent of verbal skills. These tests can be administered in pantomime and the response of the subject is entirely one of manipulation of blocks or parts.

It is interesting to note that Glaser (77,p.241) (143,p. 221) found that Jewish immigrants encountered cultural difficulties with the picture arrangement test. He has suggested, also, that anxieties handicap them particularly on the Digit Symbol test. He has argued that some of the performance items, notably those in the pictorial tests, can prove to be greater obstacles for immigrants than verbal items because of their cultural loading.

The finding of Glaser regarding pictorial tests has been confirmed by the observations of the writer at the time of testing. However, the effect of these cultural handicaps was not sufficiently great to make a difference between the results on the Wechsler-Bellevue and those on the Arthur or Cattell tests. Furthermore, this investigation has failed to provide any substantial evidence to support the claim that the performance of immigrant children on the Digit Symbol test is handicapped by their anxiety.

(21) For which of these subtests was the superiority in the performance of boys over girls most marked?

(22) On which subtest did the girls excell the boys?

The superiority in performance of boys over girls was most marked in the Object Assembly and Block Design tests. Girls excelled boys only in the Digit Symbol test. It is interesting to note from the values given in table XXI, Appendix B, that for both nationalities, there is a greater variation in the mean scores of the subtests for boys than for girls. The writer suggests the sex differential in relative performance on the subtests may be worthy of further research.

(23) For Canadians, is the average score on the Arthur Point Scale of Performance Tests (Revised Form II) higher than, lower than, or about the same as the average score for Germans?

Canadians scored higher than Germans on the average but this superiority is less on the Arthur test than on any of the other three tests (Table XII, Appendix B). In the Arthur test, there were no cultural biases obvious to the examiner. The time factor in the Seguin Form Board test did not seem to prejudice the results of either group particularly. Certainly, color-blindness would impair performance on the Stencil Design test and, possibly,

handedness and previous experience with mazes would affect the Porteus test, but in the opinion of the examiner, these factors did not operate in this experiment in a way that would invalidate comparisons.

(24) How does the average score for boys on the Arthur Point Scale of Performance Tests (Revised Form II) compare with the average score for girls?

The boys in this experiment excelled the girls on the Arthur tests. Although this male superiority was large among the Canadians, it was relatively small for the German group. (Table XI, Appendix B)

(25) On which subtest of the Arthur Point Scale did the groups perform best?

(26) On which subtest of the Arthur Point Scale did the groups tend to perform poorly?

The writer regrets that because the Arthur subtests do not contribute equally to the total score a rigorous statistical analysis is not possible. For this reason, he suggests that caution be exercised in the interpretation of the findings. An examination of the averages and ranges of the point-scores on the subtests (Table XXVI) reveals that the mean scores were 7.26 for the Knox Cube test, 7.77 on the Form Board, 8.58 for the Stencil Design test, and 6.49 for the Porteus Maze test. While it is true that on the average the subjects in this population

scored more points on the Stencil Design test than on the other tests, so also do most subjects of this age range. No valid conclusion about the relative performance on subtests can be made on the basis of mean point score values because of their differential weighting. If the scores on the subtests are converted to approximate mental age equivalents from data provided by the author (16,pp. 28-29), it appears that in relation to the established norms, the subjects herein examined performed best on the Knox Cube test and most poorly on the Porteus Maze test.

(27) For which of these subtests was the superiority in performance of Canadians over Germans most marked?

(28) For which of these subtests was the superiority in the performance of Canadians over Germans least marked?

Scores for Germans were lower than those for Canadians on all of the subtests. This difference was greatest on the Form Board test and least on the Stencil Design test.

(29) For which of these subtests was the superiority in the performance of boys over girls most marked?

(30) On which subtest did the girls excell the boys?

An examination of the point scores listed in table

XXVI reveals that the superior performance of boys is most apparent in the results of the Stencil Design test and that the girls excelled boys on the Form Board test. It should be noted that while the performance of Canadian boys was superior to that of Canadian girls, the corresponding sex difference among the Germans was smaller.

(31) Which subtest displayed the greatest discriminative power?

In terms of the range of points earned, the Seguin Form Board test appeared to have slightly more discriminative power than the other tests for the individuals in these two groups. The Seguin Form Board and the Porteus Maze tests appeared to be somewhat less discriminative among Canadians than among Germans. Such was not the case with the Knox Cube or Arthur Stencil Design tests.

(32) On which subtest of the Cattell test did the groups perform best?

The average raw-scores on the four subtests of Cattell's IPAT Test of "g": Culture-free are presented in table XXVII, Appendix B. The Conditions test appeared to be the one on which the subjects performed best. This result was surprising to the writer since the Conditions test appeared to him to be more complex than other subtests. It may be that this factor was offset by the effects of practice on the earlier subtests, greater ease

in the test situation and the generous time allotment for this subtest.

(33) On which subtest of the Cattell test did the groups perform poorly?

The Classifications test (with only 55 per cent of the responses correct) was least well done. The writer feels that the poor performance resulted from the inability of slower subjects to complete the fourteen items in four minutes and careless errors in perception.

(34) On which subtest of the Cattell test were nationality differences greatest?

If in evaluating the differences between the national groups on the four subtests, the mean scores presented in table XXVII, Appendix B are converted to percentage correct, it will be seen that the nationality averages are not greatly different. The superiority of Canadians over Germans is slightly larger in the Series test than on the others. This may be due to several factors, greater test-sophistication of Canadians is likely to apply to the Series subtest since it comes first. The nationality differences were least on the Classifications subtest.

(35) On which subtest of the Cattell test were sex differences greatest?

The performance of boys was superior to that of

girls on the Conditions and Series tests while girls clearly excelled the boys on the Classifications. In all of these, sex differences were much greater for Canadians than for the German group.

(36) How well do the test results agree with teachers' ratings of achievement?

An attempt has been made to see which test gives scores that agree most closely with grades assigned by teachers. Generally speaking, the statistical result is not satisfactory because a majority of the students were given "C" and, relatively, very few were placed in the extreme categories, "A" and "E". These grades are indicated in tables VI, VII, VIII, and IX, Appendix B.

For the experimental and the control group separately, and for each test, scores were tabulated in columns corresponding to the grades assigned by the teacher. A student rated "B" by his teacher who scored 61 on a particular test would be represented by the figure "61" appearing under column "B" on the appropriate table. An analysis of variance was completed with both groups for each test, the "F" test being applied to see if the variation due to grade was significant. These "F" values are indicated in table XXVIII, Appendix B.

With the exception of scores on the Otis test for Canadians, all of these values are significant at the five

per cent level. In all cases the means decrease progressively from the "A" to the "E" category.

These results suggest that test results and teachers' ratings agree; in only one instance is the variation in test score due to grade not sufficiently great to be significant at the five per cent level.

(37) How do the tests compare relatively in their ability to predict teacher-assigned grades for immigrant children?

(38) How do the tests compare relatively in their ability to predict teacher-assigned ratings for Canadian students?

There is an indication that test-results agree with teacher-evaluations more closely for the Germans than for Canadian students. For the Germans, the Begabungstest appeared to be the best predictor of grades. The Cattell and the Wechsler Performance tests were second and third. While the Progressive Matrices appeared to be the least satisfactory for the Germans, it was somewhat better in predicting grades for Canadian students, and in fact appeared to be as good or better than the other tests. The Otis gave results that agreed least with teachers' evaluations.

CHAPTER V

SUMMARY AND CONCLUSIONS

Summary

This investigation has compared certain mental tests for their ability to minimize cultural bias. Fifty immigrant children from Germany between the ages of ten and fifteen years enrolled in English classes for New Canadians constituted the experimental group. For control purposes, these were matched on the basis of age, sex, and raw score on the Progressive Matrices test with fifty second generation Canadians. The Begabungstest B-1 was administered to the German subjects and the Otis Self-Administering Test of Mental Ability: Intermediate Examination was given to the Canadians. The Performance Scale of the Wechsler-Bellevue Intelligence Test for Adolescents and Adults, the Arthur Point Scale of Performance Tests, and the IPAT Test of "g": Culture-free were administered to both the experimental and control groups.

Raw scores were converted into standard or normal-

ized scores with a mean of fifty and a standard deviation of ten points. Mean values of these scores were computed and comparisons were made of the results among the tests, between the groups and between sexes within the groups. Pearsonian correlation coefficients were computed between the test-results. An analysis of variance, three-way classification, was made to facilitate the study of variation in test performance due to nationality, sex, and test, and the interaction of these factors. A similar analysis of variance was performed on the weighted scores for all subjects on the five subtests of the Wechsler-Bellevue Performance Scale. In considering the variation due to tests or subtests, the Duncan multiple range test was used to segregate mean scores into homogeneous groups. Finally, an attempt was made to compare the agreement between results on the various tests with teacher evaluations.

The principal findings of this study were:

(1) Although the Progressive Matrices test (which was used for matching the individuals) would indicate that the two groups are, on the average, equal in mental ability, all other tests have rated the Canadians significantly above the German subjects.

(2) The difference in performance of the two national groups was about the same on each of the four

instruments being studied. Apparently, none of the tests is biased more than another in favor of one group. Since, for each group the verbal culture test has no bias, we may conclude that cultural factors are negligible in the other tests (the Arthur Point Scale, the IPAT Test of "g": Culture-free, and the Wechsler-Bellevue Performance Scale).

(3) The variation in test scores due to sex was greater for Canadians than for Germans, and the variation due to nationality was greater for boys than for girls.

(4) For boys, only the Arthur test gave a mean score that was about the same as the mean scores on the other tests. For girls, the mean score of the Wechsler showed the greatest agreement.

(5) The Cattell test consistently gave a significantly higher mean score than did the culture-tests.

(6) Boys performed better on both of the performance scales than on the verbal tests in this research, although the superiority of the Arthur mean score over the Culture mean score was not great enough to be significant at the .05 level. On the other hand, girls performed better on the verbal tests than on the two performance scales in this research although the superiority of the verbal test scores over those on the Wechsler was not great enough to be significant at the .05 level.

(7) Intercorrelations between the tests ranged

from .50 to .89 and all of these are significant at the .01 level. The highest degree of correlation of results (r = .89) was between the Wechsler and the Arthur tests. The Wechsler correlated best with the other tests; the Cattell and Progressive Matrices, least.

(8) In the Wechsler-Bellevue Performance Scale all groups made their highest scores on the Object Assembly test and they performed poorly on the Digit Symbol and Picture Completion test.

(9) The Picture Arrangement and Picture Completion tests favored the Canadian subjects more than the other subtests. Cultural differences were least on the Block Design and Object Assembly tests.

(10) The Object Assembly and Block Design tests were the ones on which the superiority of boys over girls was most evident. Girls excelled boys on the Digit Symbol test.

(11) There were no evidences of cultural bias noted at the time of administering the Arthur Point Scale. The difference in the mean scores for the two groups on the Arthur scale was less than that for other tests.

(12) Compared with other subtests of the Arthur point scale, the Stencil Design test minimized nationality differences; the Form Board test magnified them.

(13) While the superiority of boys over girls

was most evident on the Stencil Design test, girls excelled boys on the Form Board test.

(14) The Seguin Form Board test appeared to be somewhat more discriminative among these subjects than the other subtests of the Arthur Point Scale.

(15) In the giving of the IPAT test, the examiner saw no evidence of cultural bias. Of the four subtests the "Conditions" evoked the best response of the subjects and the "Classifications" brought forth the poorest performance. While the "Classifications" subtest appeared to minimize the differences between the two groups, for all of the subtests the differences in the nationality averages were slight.

(16) For Canadian subjects, all agreements between test results and teachers' ratings were significant at the five per cent level, except in the case of the Otis test. For the German subjects, all five tests gave results that agreed with teachers' ratings and this agreement was significant at the five per cent level. It is interesting to note that the Begabungstest B-1 was the best predictor of school grades for the German students.

Conclusions

(1) Culture-free tests can not be disregarded. In this research, three separate instruments are shown to be free from bias that would favor either of two groups of

different national and cultural background.

The difference in performance of two matched groups of subjects from different racial backgrounds was about the same on an individual "culture-free" test (the Arthur Point Scale), on a group culture-free test (the Cattell IPAT Test of "g"), and on the performance subtests of an American scale (the Wechsler-Bellevue); and these differences closely approximated the difference in their performance on the verbal culture tests. These results suggest that cultural factors have been minimized effectively and to about the same extent in these three tests. This conclusion supports earlier findings that the Arthur (168,p.499) (89,pp.419-433) and the Cattell (43,p.94) tests are relatively free from cultural influence and it suggests that in the performance scale of the Wechsler-Bellevue Intelligence Test cultural factors have been effectively reduced.

(2) In comparisons between two groups of different cultural backgrounds, their relative standings in intellectual ability can be given satisfactorily by any of these three tests.

(3) In this research, the instrument that gave the closest agreement with the culture-verbal tests regarding the relative ability of the two groups was the Cattell IPAT Test of "g". This suggests that, for inter-cultural

or inter-racial comparisons, this short test of relationships among nonverbal symbols is entirely satisfactory and it holds much promise as a fair and valid measure of intellectual ability. At the same time, it should be noted that for prediction within a group, it is less satisfactory than the individual tests. In this research, it scored individuals of both groups consistently and significantly above the rating they would receive on a verbal test.

(4) Generally speaking, the best estimate for a group of its mean score on a verbal intelligence test can be given by the Arthur Point Scale, although, in this research, this superiority was evident only with the male subjects of both groups. With girls, the Wechsler-Bellevue Performance Scale gave a somewhat closer agreement with the mean verbal score.

(5) Individual tests show significantly greater agreement in test results than do group tests. In this study the correlation between the Wechsler and Arthur tests was found to be .89.

(6) Culture-free tests give results for both Canadians and Germans that agree with teachers' evaluations. However, the verbal test was an even better predictor of school grades for German subjects.

(7) This study has revealed the possibility of using for inter-racial comparisons a scale based on American culture providing all verbal elements are eliminated. Certain subtests of the Wechsler-Bellevue Performance Scale appear to be remarkably well suited to this purpose while others appear to be culturally biased.

The fact that nationality differences on the Wechsler-Bellevue Performance Scale were greatest on the Picture Completion and Picture Arrangement tests indicates that pictorial tests fail to remove nationality differences.

(8) Since nationality differences on the Arthur Point Scale are greatest on the Form Board subtest the writer suggests that tests of psychomotor speed may fail to minimize cultural differences.

(9) In the two performance scales, nationality differences were minimized on the Stencil Design and Block Design subtests. The similarity of these tests is striking: both test the subject's ability to perceive a design, to analyze it into its parts, and to synthesize these components. The writer suggests that the greatest promise of a satisfactory solution to the problem of differential cultural background in testing lies in instruments of this type.

Suggestions for Further Research

Some of the findings that have been detailed in the previous chapter suggest areas that might well be worthy of further research.

(1) The fact that immigrant children whose scores on the Progressive Matrices are equated with those of second generation Canadians have scored consistently lower than them on the other tests (and also on the subtests of the Wechsler) seems to suggest to the writer that an analysis and a comparison need to be made of the factors being measured by the Progressive Matrices and the other tests.

(2) The analysis of variance revealed interaction between sex and nationality. Research might well be undertaken to discover the reasons underlying the greater nationality variation on these tests among boys than among girls and the greater sex variation on these tests among Canadians than among Germans.

(3) The boys in this investigation performed better on the Wechsler-Bellevue Performance test and on the Arthur Point Scale than they did on the verbal test while the girls made a better showing on the verbal test than on either of the performance scales. This may be a direct result of the nature of the samples. However, the writer suggests that a study might well be made of this implication: that boys of this age range perform better on a

performance scale than they do on a verbal test, while girls do not.

(4) There is some evidence in the results of this study that if the groups have been matched on the basis of their scores on culture-verbal tests, the Germans would have excelled the Canadians on the Progressive Matrices test and would have compared more favorably on the Cattell, Wechsler and Arthur tests. The writer suggests that this problem merits further investigation.

(5) In this research there has been an indication that girls score higher than boys on the Digit Symbol and Form Board tests and that boys excell girls on the Block Design, Object Assembly and Stencil Design tests. This sex differential in relative performance on the subtests may be worthy of further study.

(6) In view of the fact that of all the performance subtests, the pictorial and form board tests revealed the greatest difference in score between the national groups, there appears to be a need for further research into the cultural implications of tests involving pictures or psychomotor speed.

BIBLIOGRAPHY

1. Adkins, Dorothy C. and Samuel B. Lyerly. Factor analysis of reasoning tests. Chapel Hill, University of North Carolina press, 1952. 122p.

2. Allen, Robert M. and Harold Besell. Intercorrelations among group verbal and nonverbal tests of intelligence. Journal of educational research 43:394-395. 1950.

3. Alou-Bakaliar, Shah. Matritset progressiviyet b'mivhan hashvaati. (Progressive matrices in comparative testing.) Hahinukh 24:156-159. 1950/1952. (Abstracted in Psychological abstracts 27, no. 8035. 1953.)

4. Altus, William D. The comparative validities of two tests of general aptitude in an army special training center. Journal of applied psychology 30:42-44. 1946.

5. Amthauer, Rudolf. I.S.T. der intelligenz-structur test. Gottingen, Verlag fur psychologie, 1953. 38p. (Abstracted in Psychological abstracts 28, no.5197. 1954.)

6. Anastasi, Anne and Fernando A. Cordova. Some effects of bilingualism upon the intelligence test performance of Puerto Rican children in New York city. Journal of educational psychology 44:1-19. 1953.

7. Anastasi, Anne and John P. Foley. Differential psychology. Rev. ed. New York, Macmillan, 1953. 894p.

8. Arlitt, Ada H. On the need for caution in establishing race norms. Journal of applied psychology 5:179-183. 1921.

9. Arsenian, Seth. The Spearman visual perception test (part 1) with pantomime directions. British journal of educational psychology 7:287-301. 1937.

10. Arsenian, Seth. Bilingualism and mental development. New York, Teachers College, 1937. 164p. (Contributions to education no.712)

11. Arthur, Grace. A new point performance scale. Journal of applied psychology 9:390-416. 1925.

12. Arthur, Grace. An attempt to sort children with specific reading disability from other non-readers. Journal of applied psychology 11:251-263. 1927.

13. Arthur, Grace. An experience in testing Indian school children. Mental hygiene 25:188-195. 1941.

14. Arthur, Grace. A point scale of performance tests: clinical manual. 2d ed. New York, Commonwealth fund, 1943. 64p.

15. Arthur, Grace. A nonverbal test of logical thinking. Journal of consulting psychology 8:33-34. 1944.

16. Arthur, Grace. A point scale of performance tests, revised form II (Manual for administering and scoring the tests). New York, Psychological Corporation, 1947. 37p.

17. Arthur, Grace. The relative difficulty of various tests for sixty feebleminded individuals. Journal of clinical psychology 6:276-279. 1950.

18. Barke, Ethel M. A study of the comparative intelligence of children in certain bilingual and and monoglot schools in South Wales. British journal of educational psychology 3:237-250. 1933.

19. Barke, Ethel M. and D. E. Parry Williams. A further study of the comparative intelligence of children in certain bilingual and monoglot schools in South Wales. British journal of educational psychology 8:63-77. 1938.

20. Bere, May. A comparative study of the mental capacity of children of foreign parentage. New York, Columbia university, 1924. 105p. (Teachers college contributions to education no. 154)

21. Berry, Charles S. The classification by tests of intelligence of ten thousand first grade pupils. Journal of educational research 6:185-203. 1922.

22. Bessent, Trent E. A note on the validity of the Leiter international performance scale. Journal of consulting psychology 14:234. 1950.

23. Blackwood, Beatrice M. A study of mental testing in relation to anthropology. Baltimore, Williams and Wilkins, 1927. 120p. (Mental measurement monographs, serial no.4)

24. Bolton, Floyd B. Experiments with Raven's Progressive Matrices - 1938. Journal of educational research 48:629-633. 1955.

25. Britton, Joseph H. Influence of social class upon performance on the Draw-a-man test. Journal of educational psychology 45:44-51. 1954.

26. Brown, Gilbert L. Intelligence as related to nationality. Journal of educational research 5:324-327. 1922.

27. Bruner, Frank G. The hearing of primitive peoples. New York, Science press, 1908. 113p. (Archives of psychology no.11)

28. Budd, William C. Educators and culture-fair intelligence tests. Journal of educational sociology 27:333-334. 1954.

29. Burik, Theodore E. Relative roles of the learning and motor factors involved in the digit symbol test. Journal of psychology 30:33-42. 1950.

30. Buros, Oscar K. ed. The nineteen-forty mental measurements yearbook. Highland Park, Mental measurements yearbook, 1941. 674p.

31. Buros, Oscar K. ed. The fourth mental measurements yearbook. Highland Park, Gryphon press, 1953. 1163p.

32. Burt, Sir Cyril. Mental and scholastic tests. 3d ed. London, Staples press, 1947. 467p.

33. Calvin, Allen D., et al. A further investigation of the relationship between manifest anxiety and intelligence. Journal of consulting psychology 19:280-282. 1955.

34. Canada. Department of citizenship and immigration. Immigration branch. Statements for the calendar year 1954. Ottawa, 1955. 5p.

35. Carlson, Hilding B. and Norman Henderson. The intelligence of American children of Mexican parentage. Journal of abnormal and social psychology 45:544-551. 1950.

36. Cassell, Robert H. Qualitative evaluation of the progressive matrices tests. Educational and psychological measurement 9:233-241. 1949.

37. Cattell, Raymond B. Measurement versus intuition in applied psychology. Character and personality 6:114-131. 1937.

38. Cattell, Raymond B. A culture-free intelligence test I. Journal of educational psychology 31:161-179. 1940.

39. Cattell, Raymond B. Manual of directions, a culture-free test. New York, Psychological corporation, 1947. 4p.

40. Cattell, Raymond B. Classical and standard score I.Q. standardization of the I.P.A.T. culture-free intelligence scale 2. Journal of consulting psychology 15:154-159. 1951.

41. Cattell, Raymond B. A guide to mental testing. London, University of London press, 1953. 411p.

42. Cattell, Raymond B. and A. K. S. Cattell. I.P.A.T. handbook for the individual or group culture-free intelligence test, scale 2. Champaign, Institute for personality and ability testing, 1949. 9p.

43. Cattell, Raymond B., S. Norman Feingold, and Seymour B. Sarason. A culture-free intelligence test. II. Evaluation of cultural influence on test performance. Journal of educational psychology 32:81-100. 1941.

44. Cole, Luella. Psychology of adolescence. 4th ed. New York, Rinehart, 1954. 712p.

45. Cook, John Munson and Grace Arthur. Intelligence ratings for 97 Mexican children in St. Paul, Minnesota. Journal of exceptional children 18:14-15. 1951.

46. Cornell, Ethel L. and Warren W. Coxe. A performance ability scale. New York, World book, 1934. 88p.

47. Crown, Sidney. An experimental study of psychological changes following prefrontal lobotomy. Journal of general psychology 47:3-41. 1952.

48. Crown, Sidney. Psychological changes following operations on the human frontal lobe. Journal of consulting psychology 17:92-99. 1953.

49. Crown, Sidney. Psychological changes following prefrontal leucotomy: a review. Journal of mental science 97:49-83. 1951.

50. Darcy, Natalie T. The effect of bilingualism upon the measurement of the intelligence of children of preschool age. Journal of educational psychology 37:21-44. 1946.

51. Darcy, Natalie T. The performance of bilingual Puerto Rican children on verbal and on nonverbal tests of intelligence. Journal of educational research 45:499-506. 1952.

52. Darsie, Marvin L. The mental capacity of American-born Japanese children. Baltimore, Williams and Wilkins, 1926. 89p. (Comparative psychology monographs 3, no.15)

53. Dashiell, J. F. and W. D. Glenn. A re-examination of a socially composite group with Binet and with performance tests. Journal of educational psychology 16:335-340. 1925.

54. Davenport, C. B. and L. C. Crayton. Comparative traits of various races. Journal of applied psychology 7:127-134. 1923.

55. Davey, Constance M. A comparison of group verbal and pictorial tests of intelligence. British journal of psychology 17:27-48. 1926.

56. Davidson, Kenneth S. et al. A preliminary study of Negro and White differences on form I of the Wechsler-Bellevue scale. Journal of consulting psychology 14:489-492. 1950.

57. Dennis, Wayne. The performance of Hopi children on the Goodenough Draw-a-man test. Journal of comparative psychology 34:341-348. 1942.

58. Derner, Gordon F., Murray Aborn and Aaron H. Canter. The reliability of the Wechsler-Bellevue subtests and scales. Journal of consulting psychology 14:172-179. 1950.

59. Desai, Mahesh M. The relationship of the Wechsler-Bellevue verbal scale and the progressive matrices test. Journal of consulting psychology 19:60. 1955.

60. Doppelt, Jerome E. Progress in the measurement of mental abilities. Educational and psychological measurement 14:261-264. 1954.

61. DuBois, Philip H. A test standardized on Pueblo Indian children. Psychological bulletin 36:523. 1939.

62. Duncan, David B. Multiple range and multiple F tests. Biometrics 11:1-42. 1955.

63. Eells, Kenneth et al. Intelligence and cultural differences. Chicago, University of Chicago press, 1951. 388p.

64. Feingold, Gustave A. Intelligence of the first generation of immigrant groups. Journal of educational psychology 15:65-82. 1924.

65. Fils, David H. Correlation of two tests of space perception with non-language intelligence. Journal of experimental education 20:113-119. 1951.

66. Foulds, G. A. and John C. Raven. An experimental survey with the Progressive Matrices (1947). British journal of educational psychology 20:104-110. 1950.

67. Franklin, Joseph C. Discriminative value and patterns of the Wechsler-Bellevue scales in the examination of delinquent Negro boys. Educational and psychological measurement 5:71-85. 1945.

68. French, Elizabeth G. and William A. Hunt. The Navy Northwestern successive matrices test. American psychologist 4:268-269. 1949.

69. Froelich, Clifford P. Psychological testing in West Germany. Educational and psychological measurement 13:568-573. 1953.

70. Galton, Sir Francis. Hereditary genius - an inquiry into its laws and consequences. London, Macmillan, 1869. 390p.

71. Garth, Thomas R. The intelligence of full-blood Indians. Journal of applied psychology 9:382-389. 1925.

72. Garth, Thomas R. Race psychology, a study of racial mental differences. New York, McGraw-Hill, 1931. 260p.

73. Garth, Thomas R., Bert E. Lovelady, and Hale W. Smith. The intelligence and achievement of southern Negro children. School and society 32:431-435. 1930.

74. Garth, Thomas R. and Owen D. Smith. The performance of full-blood Indians on language and non-language intelligence tests. Journal of abnormal and social psychology 32:376-381. 1937.

75. Geil, George A. A clinically useful abbreviated Wechsler-Bellevue scale. Journal of psychology 20:101-108. 1945.

76. Gellerman, Saul W. and William Hays. A proposed correction for the confounded effects of cultural variation in intelligence quotients. American journal of mental deficiency 56:177-179. 1951.

77. Glaser, Nathan M. A study of the intelligence of immigrants. American psychologist 4:241. 1949.

78. Goldfarb, William. Adolescent performance in the Wechsler-Bellevue intelligence scales and the revised Stanford-Binet examination form L. Journal of educational psychology 35:503-507. 1944.

79. Goldin, Myron R. and Seymour Rothschild. Stability of intelligence quotients of metropolitan children of foreign-born parentage. Elementary school journal 42:673-676. 1942.

80. Goodenough, Florence L. The measurement of intelligence by drawings. New York, World book, 1926. 177p.

81. Goodenough, Florence L. Racial differences in the intelligence in school children. Journal of experimental psychology 9:388-397. 1926.

82. Goodenough, Florence L., Josephine G. Foster and M. J. Van Wagenen. The Minnesota preschool scale. Minneapolis, Educational test bureau, 1932. 44p.

83. Goodenough, Florence L. and Katharine M. Maurer. The mental growth of children from two to fourteen years: a study of the predictive value of the Minnesota preschool scales. Minneapolis, University of Minnesota press, 1942. 130p. (Institute of child-welfare monographs series no.20)

84. Green, Meredith W. and Josephine C. Ewert. Normative data on progressive matrices (1947). Journal of consulting psychology 19:139-142. 1955.

85. Hamilton, Mildred E. Comparison of the revised Arthur performance tests (Form II) and the 1937 Binet. Journal of consulting psychology 13:44-49. 1949.

86. Harriman, Philip L. ed. Encyclopedia of psychology. New York, Philosophical library, 1946. 897p.

87. Haught, B. F. Mental growth of the southwestern Indian. Journal of applied psychology 18:137-142. 1934.

88. Havighurst, Robert J., Minna K. Gunther, and Inez E. Pratt. Environment and the Draw-a-man test: the performance of Indian children. Journal of abnormal and social psychology 41:50-63. 1946.

89. Havighurst, Robert J. and Rhea R. Hilkevitch. The intelligence of Indian children as measured by a performance scale. Journal of abnormal and social psychology 39:419-433. 1944.

90. Healy, William. A pictorial completion test. Psychological review 21:189-203. 1914.

91. Healy, William and Grace M. Fernald. Tests for practical mental classification. Baltimore, Review publishing company, 1911. 53p. (Psychological monographs 13:no.2, whole no.54)

92. Hunter, Walter S. and Eloise Sommermeier. The relation of degree of Indian blood to score on the Otis intelligence test. Journal of comparative psychology 2:257-277. 1922.

93. Jamieson, Elmer, and Peter Sandiford. The mental capacity of Southern Ontario Indians. Journal of educational psychology 19:536-551. 1928.

94. Jones, T., C. G. Hey and W. D. Wall. A group performance test and scale of intelligence. British journal of educational psychology 22: 160-172. 1952.

95. Jones, W. R. The language handicap of Welsh speaking children. British journal of educational psychology 22:114-123. 1952.

96. Jones, W. R. The influence of reading ability in English on the intelligence test scores of Welsh speaking children. British journal of educational psychology 23:114-120. 1953.

97. Keir, Gertrude. The progressive matrices as applied to school children. British journal of psychology, statistical section 2:140-150. 1949.

98. Kent, Grace Helen. A graded series of geometrical puzzles. Journal of experimental psychology 1:40-50. 1916.

99. Klineberg, Otto. An experimental study of speed and other factors in racial differences. New York, Science press, 1928. 111p. (Archives of psychology no.93)

100. Klineberg, Otto. Race differences. New York, Harper, 1935. 367p.

101. Knox, Howard A. A scale based on the work at Ellis Island for estimating mental defect. Journal of the American medical association 62:741-747. 1914.

102. Kohs, S. C. The block design test. Journal of experimental psychology 3:357-376. 1920.

103. Leiter, Russell Graydon. The Leiter international performance scale. Santa Barbara, Santa Barbara state college press, 1940. 95p.

104. Levine, Bert and Ira Iscoe. A comparison of Raven's progressive matrices (1938) with a short form of the Wechsler-Bellevue. Journal of consulting psychology 18:10. 1954.

105. Levine, Bert and Ira Iscoe. The progressive matrices (1938), the Chicago nonverbal, and the Wechsler-Bellevue on an adolescent deaf population. Journal of clinical psychology 11: 307-308. 1955.

106. Li, Jerome C. R. Principles and methods of statistics. Corvallis, the author, 1954. 496p. (First draft)

107. Lindquist, Everet F. Statistical analysis in educational research. Boston, Houghton-Mifflin, 1940. 266p.

108. Line, W. The growth of visual perception in children. London, Cambridge university press, 1931. 148p. (British journal of psychology, monograph supplement no. 15)

109. Lorge, Irving and Seth Arsenian. A comparison of the scores on the Spearman visual perception test, part 1, administered by verbal and pantomime directions. Journal of educational psychology 29:520-522. 1938.

110. Loudon, Blanche and Grace Arthur. An application of the Fernald method to an extreme case of reading disability. Elementary school journal 40:599-606. 1940.

111. Louttit, Chauncey M. and Harvey Stackman. The relationship between Porteus Maze and Binet test performance. Journal of educational psychology 27:18-25. 1936.

112. McNemar, Quinn. Review: intelligence and cultural differences. Psychological bulletin 49:370-371. 1952.

113. MacPhee, H. M., H. F. Wright and S. B. Cummings, Jr. The performance of mentally subnormal rural southern Negroes on the verbal scale of the Bellevue intelligence examination. Journal of social psychology 25:217-229. 1947.

114. Mann, Cecil W. Mental measurements in primitive communities. Psychological bulletin 37:366-395. 1940.

115. Martin, Anthony W. and James E. Wiechers. Raven's colored progressive matrices and the Wechsler intelligence scale for children. Journal of consulting psychology 18:143-144. 1954.

116. Matarazzo, Joseph D. et al. The relationship between anxiety level and several measures of intelligence. Journal of consulting psychology 18:201-205. 1954.

117. Matarazzo, Joseph D. and Jeanne S. Phillips. Digit symbol performance as a function of increasing levels of anxiety. Journal of consulting psychology 19:131-134. 1955.

118. Matarazzo, Ruth G. The relationship of manifest anxiety to Wechsler-Bellevue subtest performance. Journal of consulting psychology 19:218. 1955.

119. Mead, Margaret. Group intelligence tests and linguistic disability among Italian children. School and society 25:465-468. 1927.

120. Mech, Edmund V. Item analysis and discriminative value of selected Wechsler-Bellevue subtests. Journal of educational research 47:241-260. 1953.

121. Mitchell, A. J. The effect of bilingualism in the measurement of intelligence. Elementary school journal 38:29-37. 1937.

122. Mursell, James L. Psychological testing. New York, Longmans Green, 1947. 449p.

123. Neff, Walter S. Socioeconomic status and intelligence: a critical survey. Psychological bulletin 35:727-757. 1938.

124. Newland, T. Ernest and William C. Lawrence. Chicago nonverbal examination results on an East Tennessee Negro population. Journal of clinical psychology 9:44-47. 1953.

125. Norsworthy, Naomi. The psychology of mentally deficient children. New York, Science press, 1906. 111p. (Archives of psychology no. 1)

126. Notcutt, Bernard. The distribution of scores on Raven's progressive matrices test. British journal of psychology, general section 40: 68-70. 1949.

127. Otis, Arthur S. Otis self-administering tests of mental ability, manual of directions and key. New York, World book, 1928. 12p.

128. Penrose, Lionel S. and John C. Raven. A new series of perceptual tests: preliminary communication. British journal of medical psychology 16:97-104. 1936.

129. Pierce-Jones, John and Fred T. Tyler. A comparison of the American council on education psychological examination and the culture-free test. Canadian journal of psychology 4:109-114. 1950.

130. Pintner, Rudolf. The standardization of Knox's cube test. Psychological review 22:377-401. 1915.

131. Pintner, Rudolf. Comparison of American and foreign children on intelligence tests. Journal of educational psychology 14:292-295. 1923.

132. Pintner, Rudolf. Intelligence testing: methods and results. New York, Holt, 1931. 555p.

133. Pintner, Rudolf and Seth Arsenian. The relation of bilingualism to verbal intelligence and school adjustment. Journal of educational research 31:255-263. 1937.

134. Pintner, Rudolf and Donald G. Paterson. A scale of performance tests. New York, Appleton, 1925. 217p.

135. Porteus, Stanley D. Temperament and mentality in maturity, sex and race. Journal of applied psychology 7:57-74. 1924.

136. Porteus, Stanley D. The maze test and mental differences. Vineland, Smith publishing company, 1933. 219p.

137. Porteus, Stanley D. Qualitative performance in the maze test. New York, Psychological corporation, 1942. 55p.

138. Porteus, Stanley D. The Porteus maze test and intelligence. Palo Alto, Pacific books, 1950. 194p.

139. Porteus, Stanley D. A survey of recent results obtained with the Porteus maze test. British journal of educational psychology 22:180-188. 1952.

140. Preston, Ralph C. Recent work in objective test construction in Germany. Educational and psychological measurement 14:381-386. 1954.

141. Rabin, Albert I. A short form of the Wechsler-Bellevue test. Journal of applied psychology 27:320-324. 1943.

142. Rabin, Albert I. The use of Wechsler-Bellevue scales with normal and abnormal persons. Psychological bulletin 42:410-422. 1945.

143. Rabin, Albert I. and Wilson H. Guertin. Research with the Wechsler-Bellevue test: 1945-1950. Psychological bulletin 48:211-248. 1951.

144. Raven, John C. The R.E.C.I. series of perceptual tests: an experimental survey. British journal of medical psychology 18:16-34. 1939.

145. Raven, John C. Standardization of progressive matrices (1938). British journal of medical psychology 19:137-150. 1941.

146. Raven, John C. The comparative assessment of intellectual ability. British journal of psychology, general section 39:12-19. 1948.

147. Raven, John C. Guide to using progressive matrices (1938). London, H. K. Lewis and company, 1952. 16p.

148. Raven, John C. and A. Waite. Experiments on physically and mentally defective children with perceptual tests. British journal of medical psychology 18:40-43. 1939.

149. Rimoldi, Horacio J. A. A note on Raven's progressive matrices test. Educational and psychological measurement 8:347-352. 1948.

150. Rimoldi, Horacio J. A., N. Cortada, and E. S. Velasco. Ensayo de tipificacion de una prueba mental (progressive matrices de Raven). Publicaciones del instituto de psicologia experimental de la universidad de Cuyo 1:81-114. 1945. (Abstracted in Psychological abstracts 20:no.930. 1946)

151. Rimoldi, Horacio J. A., et al. Tipificacion de los progressive matrices de Raven. Publicaciones del institute de psicologia experimental de la universidad de Cuyo 2:1-25. 1948. (Abstracted in Psychological abstracts 24:no.3216. 1950)

152. Rosenblum, Sidney, James E. Keller and Ned Papania. Davis-Eells ("Culture-Fair") test performance of lower-class retarded children. Journal of consulting psychology 19:51-54. 1955.

153. Russell, Roger W. The spontaneous and instructed drawings of Zuni children. Journal of comparative psychology 35:11-15. 1943.

154. Saer, D. J. The effect of bilingualism on intelligence. British journal of psychology 14:25-38. 1923.

155. Sanchez, George I. Bilingualism and mental measures. Journal of applied psychology 18:765-772. 1934.

156. Sartain, A. Q. A comparison of the new revised Stanford Binet, the Bellevue scale, and certain group tests of intelligence. Journal of social psychology 23:237-239. 1946.

157. Shotwell, Anna M. Arthur performance ratings of Mexican and American high-grade mental defectives. American journal of mental deficiency 49:445-449. 1945.

158. Sinha, Uma. A study of the reliability and validity of the progressive matrices test. Master's thesis. London, University of London, 1950. (Abstracted in British journal of educational psychology 21:238-239. 1951.)

159. Skeels, Harold M. Some Iowa studies of the mental growth of children in relation to differentials of the environment: a summary. National society for the study of education yearbook 39(2):281-308. 1940.

160. Slater, Patrick. Comment on "The comparative assessment of intellectual ability". British journal of psychology, general section 39:20-21. 1948.

161. Spearman, C. General intelligence objectively determined and measured. American journal of psychology 15:201-293. 1904.

162. Stacey, Chalmers L. and Frederick O. Carleton. The relationship between Raven's colored progressive matrices and two tests of general intelligence. Journal of clinical psychology 11: 84-85. 1955.

163. Stacey, Chalmers L. and Marie R. Gill. Relationship between Raven's colored progressive matrices and two tests of general intelligence for 172 subnormal adult subjects. Journal of clinical psychology 11:86-87. 1955.

164. Stalnaker, Elizabeth M. A study of several psychometric tests as a basis for guidance on the junior high school level. Journal of experimental education 20:41-66. 1951.

165. Stenquist, John L. and Irving Lorge. Implications of intelligence and cultural differences. Teachers college record 54:184-193. 1953.

166. Stephenson, William. Tetrad differences for verbal subtests relative to nonverbal subtests. Journal of educational psychology 22:334-350. 1931.

167. Straus, Murray A. Mental ability and cultural needs: a psychocultural interpretation of the intelligence test performance of Ceylon university entrants. American sociological review 16:371-375. 1951.

168. Tate, Miriam E. The influence of cultural factors on the Leiter international performance scale. Journal of abnormal and social psychology 47: 497-501. 1952.

169. Thomson, Godfrey H. An analysis of performance test scores of a representative group of Scottish children. London, University of London, 1940. 58p. (Publications of the Scottish council for research in education no.16)

170. Thorndike, Robert L. Community variables as predictors of intelligence and academic achievement. Journal of educational psychology 42: 321-338. 1951.

171. Thurstone, Louis L. Primary mental abilities, 1938. Chicago, University of Chicago press, 1938. 121p. (Psychometric monograph serial no.1)

172. Thurstone, Louis L. and Thelma G. Thurstone. Factorial studies of intelligence. Chicago, University of Chicago press, 1941. 94p. (Psychometric monograph serial no.2)

173. Tilton, John R. A survey of the reliability, validity, and usefulness of the Cattell culture-free test. Persona 1:17-19. 1949.

174. Tilton, John W. The intercorrelations between measures of school learning. Journal of psychology 35:169-179. 1953.

175. Tizard, John. Porteus maze test and intelligence: a critical survey. British journal of educational psychology 21:172-185. 1951.

176. Turner, G. H. and D. J. Penfold. The scholastic aptitude of the Indian children of the Caradoc reserve. Canadian journal of psychology 6: 31-44. 1952.

177. Tyler, Fred. Comments on the correlational analysis reported in "Intelligence and cultural differences". Journal of educational psychology 44:288-295. 1953.

178. Vernon, Philip E. Intelligence test sophistication. British journal of educational psychology 8: 237-244. 1938.

179. Vernon, Philip E. and John B. Parry. Personnel selection in the British forces. London, University of London press, 1949. 324p.

180. Warner, William Lloyd, Marchia Meeker, and Kenneth Eells. Social class in America: a manual of procedure for the measurement of social status. Chicago, Science research associates, 1949. 274p.

181. Watson, Robert I. The use of the Wechsler-Bellevue scales: a supplement. Psychological bulletin 43:61-68. 1946.

182. Wechsler, David. The measurement of adult intelligence. Rev.ed. Baltimore, Williams and Wilkins, 1944. 258p.

183. Weisenberg, Theodore H., Anne Roe and Katharine E. McBride. Adult intelligence. New York, Commonwealth fund, 1936. 155p.

184. Woodsworth, Robert S. Racial differences in mental traits. Science 31:171-186. 1910.

185. Yerkes, Robert M. ed. Psychological examining in the United States army. Washington, Government printing office, 1921. 890p. (Memoirs national academy of sciences 15.)

186. Yoakum, Clarence S. and Robert M. Yerkes. eds. Army mental tests. New York, Holt, 1920. 303p.

187. Young, Kimball. Mental differences in certain immigrant groups. Eugene, University of Oregon, 1922. 103p. (University of Oregon publication no.11)

APPENDIX A

APPENDIX A

The Use of a New Multiple Range Test

The common practice for testing the homogeneity of a set of means in an analysis of variance is to use an F (or z) test. This procedure has several special desirable properties for testing the homogeneity hypothesis that the 'n' population means concerned are equal. Duncan (62,pp.1-42) points out that an F test alone generally falls short of satisfying all of the practical requirements involved. When it rejects the homogeneity hypothesis, it gives no decision as to which of the differences among the treatment means may be considered significant and which may not. Several test procedures have been devised to answer this problem.

The Duncan test used herein is a multiple range test that combines the features considered best from previously proposed tests. From a table of significant studentized ranges for a 5 per cent level test, values are extracted appropriate to the number of means to be tested and the number of degrees of freedom. The significant studentized ranges are then multiplied by the standard error to form what may be called the shortest significant ranges. If the difference between any two means is less than the shortest significant range, then means may be

considered to be similar; if the difference is equal to or greater than this amount, the means may be considered to be different.

To illustrate the use of this test, the mean scores for the boys are arranged in ranked order, decreasing from left to right:

Wechsler	Cattell	Arthur	Culture
56.50	55.65	54.54	53.52

Then the table of significant studentized ranges for a 5 per cent level test (62,p.3) is entered at the appropriate row for 288 degrees of freedom, (in this particular table the appropriate row is for ∞, an infinite number of degrees of freedom) and from this row the significant-studentized ranges are extracted for samples of sizes 2, 3 and 4 means. The values obtained in this way are 2.77, 2.92 and 3.02. These significant studentized ranges are then multiplied by the standard error ($\sqrt{\frac{s^2}{n}}$), (where 's^2' is the variance and 'n' is the number of subjects in the sample) to form what Duncan calls the "shortest significant ranges". From table 4, $s^2 = 23.75$ and for boys $n = 46$; $\sqrt{\frac{s^2}{n}} = .71854$. When this factor is multiplied by each of the values above (2.77, 2.92 and 3.02) we find that the shortest significant ranges for 2, 3, or 4 means in this case are 1.99, 2.10 and 2.17, respectively.

The differences are tested in the following order:

the largest minus the smallest, the largest minus the second smallest, up to the largest minus the second largest; then the second largest minus the smallest, the second largest minus the second smallest, and so on, finishing with the second smallest minus the smallest.

In this case the Wechsler score exceeds the Culture score by 2.98. Since this is greater than 2.17, the shortest significant range for four means, we may conclude that for boys the Wechsler mean score and the Culture test mean score are different from one another.

The difference between the Wechsler and the Arthur means (1.96) is less than the shortest-significant-range of three means (2.10) and hence it is not significant. The Wechsler and the Arthur means for boys are considered to be similar.

Duncan states that "no difference between two means can be declared significant if the two means concerned are both contained within a subset of the means which has a non-significant range" (62,p.6). On this basis the Cattell mean which lies between the Wechsler and Arthur means must be considered similar to both of them. The three means are homogeneous.

The difference between the means for boys on the Cattell and the Culture tests (2.13), since it exceeds the shortest-significant range (2.10) for three means, is

significant. The difference between the means on the Arthur and Culture tests (1.02) is smaller than the shortest-significant range (1.99) for two means and therefore it is not significant.

These results are recorded by drawing a line under means that are not significantly different from one another; thus:

Wechsler	Cattell	Arthur	Culture
56.50	55.65	54.54	53.52

Any two means not underscored by the same line are significantly different.

APPENDIX B

TABLE I

IMMIGRATION TO CANADA, BY RACIAL ORIGIN

	1953	1954
German	35,015	29,845
English	28,325	26,714
Italian	24,293	24,595
Dutch	20,472	16,340
Scottish	10,344	10,480
United States	9,379	10,110
Irish	7,562	6,438
Austrian	3,574	3,841
Greek	2,059	2,892
French	3,136	2,813
Polish	3,176	2,274
Chinese	1,929	1,950
Others	19,604	15,935
Total	168,868	154,227

TABLE II

IMMIGRATION TO CANADA
SHOWING DESTINATION FOR THE CALENDAR YEAR, 1954

Destination	No. of Persons
Newfoundland	524
Nova Scotia	2,207
New Brunswick	1,011
Prince Edward Island	107
Quebec	28,419
Ontario	83,029
Manitoba	9,219
Saskatchewan	4,125
Alberta	13,294
British Columbia	12,197
Yukon and Northwest Territories	95
Total	154,227

Data supplied by Canadian Department of Citizenship and Immigration. (34,pp.1-3)

TABLE III

IMMIGRATION TO CANADA BY AGE GROUPS AND SEX CALENDAR YEAR 1954

Age groups	Totals	Males	Females
0 - 14	33,098	17,222	15,876
15 - 19	11,307	6,475	4,832
20 - 24	29,033	16,444	12,589
25 - 29	29,072	17,080	11,992
30 - 39	29,818	16,708	13,110
40 - 49	13,718	7,393	6,325
50 - 59	5,159	2,180	2,979
60 and over	3,022	1,029	1,993
Totals	154,227	84,531	69,696

TABLE IV

COMPARISONS OF CHRONOLOGICAL AGES (AT THE TIME OF TESTING) AND PROGRESSIVE MATRICES SCORES OF 23 MATCHED PAIRS OF MALE SUBJECTS

	Chronological Age in Years and Months			Progressive Matrices Scores		
Pair No.	Germans	Canadians	Dif.* in Months	Germans	Canadians	Difference*
1	12-4	12-4	0	38	39	1
2	15-0	14-11	-1	54	54	0
3	13-7	13-8	1	40	39	-1
4	10-7	10-4	-3	45	46	1
5	10-4	10-7	3	38	37	-1
6	13-7	13-5	-2	47	48	1
7	10-2	10-5	3	34	35	1
8	14-1	14-3	2	47	46	-1
9	10-8	10-11	3	51	52	1
10	12-9	12-10	1	42	42	0
11	15-10	15-9	-1	48	47	-1
12	11-5	11-3	-2	25	25	0
13	14-2	14-2	0	39	40	1
14	13-0	13-1	1	50	50	0
15	12-8	12-9	1	50	51	1
16	15-6	15-6	0	43	43	0
17	12-9	12-10	1	38	38	0
18	12-1	12-1	0	45	42	-3
19	11-7	11-9	2	33	32	-1
20	14-3	14-3	0	48	46	-2
21	11-11	11-10	-1	48	51	3
22	11-9	11-7	-2	42	43	1
23	10-4	10-3	-1	42	43	1
Mean Algebraic Difference			0.2 mo.			0.1
Mean Arithmetical Difference			1.3 mo.			1.0

(*Differences are expressed in terms of Canadians less Germans.)

TABLE V

COMPARISONS OF CHRONOLOGICAL AGES (AT THE TIME OF TESTING) AND PROGRESSIVE MATRICES SCORES OF 27 MATCHED PAIRS OF FEMALE SUBJECTS

Chronological Age in Years and Months				Progressive Matrices Scores		
Pair No.	Germans	Canadians	Dif.* in Months	Germans	Canadians	Difference*
1	12-11	13-1	2	31	32	1
2	11-7	11-10	3	25	25	0
3	14-2	14-2	0	34	33	-1
4	13-3	13-2	-1	46	47	1
5	15-6	15-5	-1	36	37	1
6	14-2	14-4	2	50	49	-1
7	13-9	13-9	0	48	48	0
8	10-4	10-6	2	38	36	-2
9	15-1	15-0	-1	45	46	1
10	12-5	12-6	1	37	38	1
11	10-10	10-8	-2	45	44	-1
12	12-8	12-9	1	40	41	1
13	14-1	14-2	1	37	37	0
14	11-10	12-0	2	39	39	0
15	13-9	13-11	2	50	51	1
16	10-8	10-9	1	40	39	-1
17	11-6	11-8	2	48	50	2
18	13-10	13-10	0	46	45	-1
19	14-10	15-0	2	50	49	-1
20	12-5	12-6	1	51	50	-1
21	10-4	10-10	6	14	19	5
22	10-11	11-0	1	41	39	-2
23	15-4	15-8	4	51	52	1
24	11-8	11-6	-2	32	33	1
25	14-0	13-10	-2	31	30	-1
26	11-2	11-6	4	43	42	-1
27	11-7	11-8	1	46	44	-2
Mean Algebraic Difference			1.1 mo.			0.0
Mean Arithmetical Difference			1.7 mo.			1.2

(*Differences are expressed in terms of Canadians less Germans.)

TABLE VI

STANDARD-SCORES FOR 23GERMAN BOYS ON THE BEGABUNGSTEST, WECHSLER-BELLEVUE, ARTHUR AND CATTELL TESTS, THEIR NORMALIZED SCORES ON THE PROGRESSIVE MATRICES TEST AND TEACHERS' RATINGS OF THEIR ACHIEVEMENT

No.	Bega-bungs test	Wechsler Bellevue	Arthur	Cattell	Progres-sive Matrices	Teachers' Ratings
1	45	39	35	41	48	C
2	62	66	63	60	67	A
3	45	44	43	50	46	C
4	56	58	57	59	63	B
5	45	57	42	52	56	C
6	61	48	40	56	55	C
7	41	53	54	48	53	C
8	55	59	58	58	55	B
9	52	64	64	61	68	C
10	39	44	37	39	50	C
11	30	44	43	33	56	C
12	31	43	39	45	42	D
13	40	44	42	44	45	C
14	48	54	53	56	63	C
15	60	66	65	61	63	A
16	47	54	55	52	49	C
17	57	48	45	50	47	C
18	50	63	60	55	56	B
19	46	51	52	47	47	C
20	51	51	55	50	56	C
21	55	51	47	59	61	C
22	55	61	54	55	54	C
23	61	61	59	53	60	C

TABLE VII

STANDARD-SCORES FOR 27 GERMAN GIRLS ON THE BEGABUNGSTEST, WECHSLER-BELLEVUE, ARTHUR AND CATTELL TESTS, THEIR NORMALIZED SCORES ON THE PROGRESSIVE MATRICES TEST AND TEACHERS' RATINGS OF THEIR ACHIEVEMENT

No.	Bega-bungs test	Wechsler Bellevue	Arthur	Cattell	Progres-sive Matrices	Teachers' Ratings
1	39	39	45	44	41	C
2	51	56	50	50	43	C
3	51	43	37	46	42	C
4	51	47	45	47	54	C
5	45	35	37	50	42	C
6	57	49	51	56	59	B
7	48	48	62	54	57	C
8	51	54	50	50	56	C
9	52	44	40	52	52	C
10	51	50	43	56	47	C
11	56	55	57	67	61	B
12	45	52	45	59	48	C
13	47	44	49	50	43	C
14	47	50	48	47	51	B
15	53	60	51	58	59	B
16	49	40	41	46	57	C
17	65	64	59	53	61	A
18	49	49	60	57	53	C
19	59	57	55	56	59	B
20	61	64	64	64	64	A
21	41	36	44	44	37	C
22	49	60	57	56	57	D
23	48	58	53	52	60	C
24	44	46	48	52	45	C
25	30	36	42	30	39	E
26	47	56	42	53	58	C
27	57	50	55	61	59	B

TABLE VIII

STANDARD-SCORES FOR 23 CANADIAN BOYS ON THE OTIS, WECHSLER-BELLEVUE, ARTHUR AND CATTELL TESTS, THEIR NORMALIZED SCORES ON THE PROGRESSIVE MATRICES TEST, AND TEACHERS' RATINGS OF THEIR ACHIEVEMENT

No.	Otis	Wechsler Bellevue	Arthur	Cattell	Progressive Matrices	Ratings
1	60	61	56	64	49	B
2	66	65	64	62	67	B
3	52	54	52	61	45	C
4	66	68	80	75	65	C
5	57	64	65	62	54	C
6	58	61	60	62	57	B
7	62	68	57	63	52	C
8	53	68	53	64	53	C
9	56	63	68	62	68	B
10	59	62	61	68	50	C
11	50	54	49	50	54	C
12	38	54	50	47	42	D
13	48	46	41	60	46	C
14	69	66	69	65	63	B
15	71	69	65	59	64	B
16	42	46	48	56	49	C
17	49	46	45	46	47	C
18	65	59	58	51	53	C
19	53	49	48	57	45	C
20	55	46	43	48	53	C
21	70	70	78	67	66	C
22	60	71	78	65	56	B
23	71	66	59	62	62	B

TABLE IX

STANDARD-SCORES FOR 27 CANADIAN GIRLS ON THE OTIS, WECHSLER-BELLEVUE, ARTHUR AND CATTELL TESTS, THEIR NORMALIZED-SCORES ON THE PROGRESSIVE MATRICES TEST AND TEACHERS' RATINGS OF THEIR ACHIEVEMENT

No.	Otis	Wechsler Bellevue	Arthur	Cattell	Progressive Matrices	Ratings
1	47	47	43	43	42	C
2	45	42	39	64	40	D
3	54	50	57	50	41	C
4	61	61	53	57	56	C
5	39	43	44	50	43	D
6	61	68	66	62	58	B
7	51	48	43	55	57	C
8	50	61	52	61	53	C
9	55	53	56	58	53	B
10	48	59	52	50	48	C
11	70	61	56	59	61	B
12	65	62	50	63	49	B
13	49	35	40	58	43	C
14	48	51	50	53	50	C
15	56	51	57	62	61	C
16	60	63	62	60	56	C
17	68	62	58	65	64	C
18	60	61	52	68	52	B
19	57	57	46	52	58	B
20	72	64	59	73	63	A
21	64	49	48	39	40	C
22	65	61	52	64	54	C
23	52	64	58	62	63	C
24	51	50	60	47	47	C
25	40	47	38	36	39	E
26	55	51	53	61	56	C
27	55	58	58	64	57	C

TABLE X

ANALYSIS OF VARIANCE OF SCORES ON FOUR MENTAL TESTS FOR 50 GERMAN AND 50 CANADIAN SUBJECTS, THREE-WAY CLASSIFICATION FOR NATIONALITY, SEX, AND TEST

Variation due to:	Sum of Squares	Degrees of Freedom	Mean Square	F
Nationality	3,708.81	1	3,708.81	156.16*
Sex	648.29	1	648.29	27.30*
Nat. x Sex	283.11	1	283.11	11.92*
Student	19,853.54	96	206.81	8.71*
Test	421.01	3	140.34	5.91*
Test x nat.	50.01	3	16.67	0.70
Test x sex	240.55	3	80.18	3.38*
Test x nat x sex	47.12	3	15.71	0.66
Test x student	6,839.31	288	23.75	
Total	32,091.75	399		

*Significant at .05 level.

TABLE XI

AVERAGE SCORES ON FOUR MENTAL TESTS FOR THE 23 BOYS AND THE 27 GIRLS OF THE GERMAN AND CANADIAN GROUPS

Students	'Culture' Verbal Test	Wechsler Bellevue Perf.	Arthur Point Scale	Cattell	Students' Averages
German Boys (N=23)	49.22	53.17	50.52	51.48	51.10
German Girls (N=27)	49.74	49.70	49.26	52.22	50.23
Canadian Boys (N=23)	57.83	59.83	58.57	59.83	59.01
Canadian Girls (N=27)	55.48	54.78	51.93	56.89	54.77
Test Averages	53.03	54.20	52.41	55.06	53.68

TABLE XII

AVERAGE SCORES ON FOUR MENTAL TESTS FOR THE GERMAN AND THE CANADIAN GROUPS (SEX DISAPPEARS)

Students	'Culture' Verbal Test	Wechsler Bellevue Perf.	Arthur Point Scale	Cattell	Nationalities' Averages
German (N=50)	49.50	51.30	49.84	51.88	50.63
Canadians (N=50)	56.56	57.10	54.98	58.24	56.72
Test Averages (N=100)	53.03	54.20	52.41	55.06	53.68

TABLE XIII

AVERAGE SCORES FOR THE 23 BOYS AND THE 27 GIRLS
OF BOTH GERMAN AND CANADIAN GROUPS
(TESTS DISAPPEAR)

Students	Boys	Girls	Nationalities' Averages
Germans (N=50)	51.10 (N=23)	50.23 (N=27)	50.63 (N=50)
Canadians (N=50)	59.01 (N=23)	54.77 (N=27)	56.72 (N=50)
Sex Averages (N=100)	55.05 (N=46)	52.50 (N=54)	53.68 (N=100)

TABLE XIV

AVERAGE SCORES FOR 46 BOYS AND 54 GIRLS
OF BOTH GROUPS ON FOUR MENTAL TESTS
(NATIONALITY DISAPPEARS)

Students	'Culture' Verbal Test	Wechsler Bellevue Perf.	Arthur Point Scale	Cattell	Sex Averages
Boys (N=46)	53.52	56.50	54.54	55.65	55.05
Girls (N=54)	52.61	52.24	50.59	54.56	52.50
Test Averages (N= 100)	53.03	54.20	52.41	55.06	53.68

TABLE XV

DIFFERENCES* IN MEAN SCORES ON FOUR TESTS BETWEEN GERMANS AND CANADIANS FOR BOTH SEXES

Students	'Culture' Verbal Test	Wechsler Bellevue Perf.	Arthur Point Scale	Cattell	Average
Boys	8.61	6.66	8.05	8.35	7.91
Girls	5.74	5.08	2.67	4.67	4.54
Boys & Girls	7.06	5.80	5.14	6.36	6.09

TABLE XVI

WEIGHTED-SCORES ON FIVE PERFORMANCE SUBTESTS OF THE WECHSLER-BELLEVUE TEST FOR 23 GERMAN BOYS

No.	Picture Arrangement	Picture Completion	Block Design	Object Assembly	Digit Symbol
1	6	3	7	7	7
2	11	14	14	14	11
3	9	7	7	10	7
4	6	6	8	12	6
5	7	8	7	10	5
6	9	6	10	12	7
7	6	7	5	9	5
8	13	8	13	13	8
9	7	10	11	10	6
10	11	7	7	7	5
11	6	10	10	8	11
12	7	6	5	7	4
13	11	8	9	9	6
14	6	9	14	12	7
15	14	13	11	12	8
16	7	10	13	14	10
17	7	10	9	10	6
18	12	9	12	11	8
19	4	9	8	11	6
20	11	9	11	11	7
21	6	9	7	11	7
22	7	8	10	14	9
23	6	7	10	10	7

TABLE XVII

WEIGHTED-SCORES ON FIVE PERFORMANCE SUBTESTS OF THE WECHSLER-BELLEVUE TEST FOR 27 GERMAN GIRLS

No.	Picture Arrangement	Picture Completion	Block Design	Object Assembly	Digit Symbol
1	7	3	6	9	9
2	7	12	7	11	6
3	10	8	7	7	9
4	11	6	10	9	7
5	7	4	7	11	7
6	3	10	12	12	10
7	7	8	9	10	10
8	10	3	6	9	5
9	4	9	10	13	7
10	7	8	8	10	8
11	7	9	9	6	7
12	13	4	11	11	6
13	7	9	7	10	9
14	6	7	8	12	6
15	14	10	12	11	7
16	7	3	4	4	5
17	11	8	13	10	7
18	7	10	11	8	10
19	11	12	11	10	11
20	9	12	10	15	8
21	3	1	4	4	6
22	6	10	7	14	6
23	9	12	13	12	11
24	6	2	8	11	7
25	6	8	6	8	7
26	8	9	5	12	6
27	6	7	7	9	8

TABLE XVIII

WEIGHTED-SCORES ON FIVE PERFORMANCE SUBTESTS OF THE WECHSLER-BELLEVUE TEST FOR 23 CANADIAN BOYS

No.	Picture Arrangement	Picture Completion	Block Design	Object Assembly	Digit Symbol
1	12	10	10	12	9
2	10	13	13	14	13
3	11	9	11	12	8
4	7	9	12	12	7
5	12	7	10	10	6
6	14	13	10	13	7
7	11	9	7	12	9
8	17	13	10	15	10
9	8	7	12	11	8
10	10	12	11	13	9
11	9	10	13	15	7
12	13	8	5	8	6
13	8	9	9	12	7
14	11	10	14	14	10
15	13	12	13	15	9
16	9	12	10	7	8
17	4	10	6	12	8
18	9	10	10	11	9
19	5	6	6	11	10
20	11	9	8	10	7
21	15	9	14	12	8
22	14	12	12	12	8
23	8	9	10	12	7

TABLE XIX

WEIGHTED-SCORES ON FIVE PERFORMANCE SUBTESTS OF THE WECHSLER-BELLEVUE TEST FOR 27 CANADIAN GIRLS

No.	Picture Arrangement	Picture Completion	Block Design	Object Assembly	Digit Symbol
1	8	8	9	10	7
2	4	9	3	9	7
3	9	9	6	13	11
4	13	10	10	13	10
5	7	9	9	7	12
6	13	9	15	14	14
7	11	6	8	11	9
8	8	8	7	12	6
9	8	9	10	12	13
10	13	10	11	10	7
11	11	9	4	12	7
12	13	9	10	12	11
13	5	10	4	4	11
14	7	10	7	10	7
15	9	9	10	9	12
16	11	8	9	10	7
17	14	9	8	9	10
18	8	10	15	14	10
19	11	13	10	12	10
20	14	10	12	10	9
21	7	4	7	7	8
22	8	10	9	12	7
23	13	13	14	11	14
24	9	8	5	6	9
25	11	10	8	7	9
26	14	8	5	6	6
27	10	10	10	8	8

TABLE XX

A THREE-WAY ANALYSIS OF VARIANCE OF THE SCORES ON FIVE PERFORMANCE SUBTESTS OF THE WECHSLER-BELLEVUE INTELLIGENCE TEST

Variation Due to:	Sum of Squares	Degrees of Freedom	Mean Square	F
Nationality	206.08	1	206.08	50.76*
Sex	49.01	1	49.01	12.07*
Nat. x sex	1.90	1	1.90	0.47
Students	1,416.07	96	14.75	3.63*
Test	332.73	4	83.18	20.49*
Test x nat.	50.63	4	12.66	3.12*
Test x sex	72.71	4	18.18	4.48*
Test x nat. x sex	11.65	4	2.91	.72
Test x student	1,558.28	384	4.06	
Total	3,699.06	499		

*Significant at .05 level.

TABLE XXI

MEAN SCORES ON EACH OF THE SUBTESTS OF WECHSLER PERFORMANCE SCALE MADE BY BOYS AND GIRLS OF THE GERMAN AND CANADIAN GROUPS

Students	Picture Arrangement	Picture Completion	Block Design	Object Assembly	Digit Symbol	Test Averages
German Boys (N=23)	8.22	8.39	9.48	10.61	7.09	8.76
German Girls (N=27)	7.74	7.56	8.44	9.93	7.59	8.25
Canadian Boys (N=23)	10.48	9.91	10.26	11.96	8.26	10.17
Canadian Girls (N=27)	9.96	9.15	8.70	10.00	9.30	9.42
Subtest Averages (N=100)	9.08	8.72	9.17	10.57	8.09	9.13

TABLE XXII

MEAN SCORES FOR 46 BOYS AND 54 GIRLS OF BOTH GROUPS ON EACH OF THE SUBTESTS OF THE WECHSLER PERFORMANCE SCALE. (NATIONALITY DISAPPEARS)

Students	Picture Arrangement	Picture Completion	Block Design	Object Assembly	Digit Symbol	Test Averages
Boys (N=46)	9.35	9.15	9.87	11.28	7.67	9.47
Girls (N=54)	8.85	8.35	8.57	9.96	8.44	8.84
Totals (N=100)	9.08	8.72	9.17	10.57	8.09	9.13

TABLE XXIII

MEAN SCORES FOR THE GERMAN AND THE CANADIAN GROUPS ON EACH OF THE SUBTESTS OF THE WECHSLER PERFORMANCE SCALE. (SEX DISAPPEARS)

Students	Picture Arrange-ment	Picture com-pletion	Block Design	Object Assem-bly	Digit Symbol	Test Averages
German (N=50)	7.96	7.94	8.92	10.24	7.36	8.48
Canadians (N=50)	10.20	9.50	9.42	10.90	8.82	9.77
Subtest Averages	9.08	8.72	9.17	10.57	8.09	9.13

TABLE XXIV

MEAN SCORES ON THE WECHSLER PERFORMANCE SCALE FOR BOYS AND GIRLS OF BOTH GERMAN AND CANADIAN GROUPS (TESTS DISAPPEAR)

Students	Boys	Girls	Nationalities' Averages
Germans	8.76 (N=23)	8.25 (N=27)	8.48 (N=50)
Canadians	10.17 (N=23)	9.42 (N=27)	9.77 (N=50)
Sex Averages	9.47 (N=46)	8.84 (N=54)	9.13 (N=100)

TABLE XXV

Between	And	For	Coefficient*
Culture	Wechsler	Boys	.75
		Girls	.71
Culture	Arthur	Boys	.69
		Girls	.57
Culture	Cattell	Boys	.75
		Girls	.64
Culture	Matrices	Boys	.60
		Girls	.62
Wechsler	Arthur	Boys	.89
		Girls	.71
Wechsler	Cattell	Boys	.79
		Girls	.61
Wechsler	Matrices	Boys	.66
		Girls	.69
Arthur	Cattell	Boys	.76
		Girls	.53
Arthur	Matrices	Boys	.67
		Girls	.63
Cattell	Matrices	Boys	.50
		Girls	.64

*All of these coefficients are significant at the .01 level (107,p.212).

TABLE XXVI

AVERAGES AND RANGES OF POINT-SCORES ON FOUR ARTHUR SUBTESTS FOR 100 SUBJECTS

	Knox Cube	Form Board	Stencil Design	Porteus Maze
German Boys (N=23)	6.93* (4.70 - 8.60)	7.01* (3.18 - 9.93	8.78* (6.20 - 10.76)	6.24* (3.34 - 8.39)
German Girls (N=27)	6.90 (4.28 - 9.00)	7.68 (4.73 - 10.73)	8.21 (5.80 - 11.16)	6.12 (3.68 - 8.89)
Canadian Boys (N=23)	7.86 (5.28 - 10.20)	8.23 (6.73 - 9.93)	9.15 (6.93 - 11.56)	6.97 (5.39 - 8.39)
Canadian Girls (N=27)	7.38 (4.70 - 10.20)	8.11 (6.06 - 9.93)	8.30 (6.20 - 10.76)	6.61 (5.14 - 8.39)

*Average. Ranges are indicated in Parentheses.

TABLE XXVII

AVERAGE RAW-SCORES IN FOUR SUBTESTS OF CATTELL'S IPAT TEST OF "g" CULTURE-FREE FOR 23 BOYS AND 27 GIRLS OF BOTH GERMAN AND CANADIAN GROUPS

	Series (12 items)	Classifications (14 items)	Matrices (12 items)	Conditions (8 items)	Totals (46 items)
German Boys	8.00	7.13	7.73	6.13	28.99
German Girls	7.89	7.26	8.41	5.96	29.52
Canadian Boys	9.70	7.70	9.35	7.30	34.05
Canadian Girls	8.93	8.56	8.85	6.37	32.71

TABLE XXVIII

ANALYSIS OF VARIANCE 'F' VALUES FOR VARIATION DUE TO TEACHER-ASSIGNED GRADES IN THE SCORES ON FIVE MENTAL TESTS FOR 50 GERMAN AND 50 CANADIAN SUBJECTS

Students	'Culture' Verbal Test	Wechsler Bellevue Perf.	Arthur Point Scale	Cattell	Progressive Matrices
Germans	10.45*	7.30*	5.88*	8.02*	5.73*
Canadians	1.08	4.71*	3.37*	4.67*	6.22*

*Significant at .05 level.