INTELLIGENCE TESTS

F. Kuhlmann

Division of Research, Board of Control

The main object of this paper is to answer a number of miscellaneous questions frequently asked by laymen and professionals in other fields. The nature of intelligence tests and the results they give will be discussed only for the purpose of helping to answer these questions. It may seem to some that after a quarter of a century of intelligence testing this would be unnecessary. But psychology is not in the same situation as are most of the other sciences and professions. If an engineer tells the public that a bridge is unsafe beyond a certain load his verdict is seldom questioned. If a doctor says that a certain water supply is unfit for drinking few people will drink it. In hundreds of ways we daily put our trust and confidence in the judgments of other people, conscious of the fact that they are more expert than we in the special fields outside our own. The psychologist does not command the same degree of confidence. There are probably a number of reasons for it. Two may be mentioned as outstanding. The first is that psychology more directly affects everyone in his everyday life than does any other profession. We are not daily and hourly concerned with engineers and doctors, but we are concerned every minute with what we are, mentally, emotionally, and spritually. This fact has made everyone his own psychologist, and that is the second reason why the professionally trained psychologist of today has difficulty in gaining the same respect that is enjoyed in some other professions. Man has judged his fellow men as to intelligence from his very origin. Every layman who has never heard of intelligence tests does so daily, habitually, and as a matter of necessity in making his own adjustments and in determining his reactions towards his associates. Who then is Binet, or Thorndike, or Terman to tell him that he is unable to do this, and that the psychologist alone with his new and special device that he calls "intelligence tests" can do it. correctly?

Can Intelligence Re Measured? Instead of answering this question at this point, which is really the basic part of the whole issue, let us turn to another that reflects the exactly opposite point of view. This question asks how such an intangible and (indefinable thing as intelligence can be tested or measured at all. Many capable men, including a few psychologists, are still sincerely and earnestly asking this question. Let us recognize the fact that there is no universally or even very commonly accepted definition of intelligence. No group or scale of intelligence tests ever devised has consistently followed any definite or particular definition. There are two parts to the answer. The first is that we do not measure intelligence. The term itself is only a convenient abstraction. We measure only what intelligence does, the product, not the thing itself. To do this does not require any precise definition, or even any definition at all. Physical sciences do the same. We measure what electricity does. We do not measure electricity. As noted by Dr. P. R. Hey of the National Bureau of Standards in a recent article on "What is Electricity?" in the July 1935 number of the Scientific Monthly, we have not defined it. "I trust," he says, "there is no one in this audience so optimistic as to suppose that because I have asked this question I am going to answer it." Another equally apt illustration is gravity. In weighing things we measure what gravity does, not gravity itself. And no one has yet defined gravity. The psychologist is not doing anything unusual in this respect when he proposes to measure intelligence without being able to define it. One might, indeed, ask the questioner how he measures intelligence without defining it. For he could hardly deny that as a matter of fact he does make distinctions between different people on the basis of different amounts of intelligence. Thus it becomes clear that the person who asks how the. psychologist can measure intelligence without being able to define it has the same idea and attitude as the one who claims that his own methods which he has always used are as good or better than intelligence tests. Both deny that the intelligence test is any improvement over older methods that have always been practiced. What, those older methods are will be considered briefly after considering the nature of the intelligence test.

Confidence In Intelligence Tests: Many of those who do not subscribe to intelligence tests find it easy to believe that most other people hold the same opinion. The continued reiteration of statements to this effect has caused confusion and doubt in the minds of many who do not have the means of learning the facts for themselves. We are told that "people do not believe in intelligence tests;" that, the "average man" has no faith in the psychologists; that judges do not accept test results as of value; and so on. These and similar statements are in pu't unfounded generalisations from exceptional instances, and in part meaningless ambiguities. For example, What "people" are referred to, and who is the "average man"? Undoubtedly the majority of adults in any community have no opinion or belief whatsoever about intelligence tests. They have had no occasion to give any thought to them. If we limited ourselves to people who have a right to any opinion on the matter, we would find quite the opposite sentiment. The popularity of intelligence tests and the faith that informed people have in them can be gauged somewhat by observing how frequently they are used and by whom. On this matter wc may note that there are several hundred mental clinics in this country that employ psychologists for the purpose of giving intelligence tests. They are used routinely in state institutions of most states. All the larger school systems of the country have psychologists on the school staff, and many more use teachers in that capacity. The Division of Mental Deficiency of the State Department of Mental Diseases in Massachusetts makes over 8,000 individual mental test examinations a year. Similar numbers are examined yearly by the State Bureaus of Juvenile Research in Illinois, Ohio, and California. Our own Division of Research makes about 5,000 examinations a year, and, like the others undoubtedly, never kept. up even approximately with demands. The number examined in the public, parochial and private schools is undeterminable, but is unquestionably many times the number examined by state departments and private clinics combined. All this concerns only examinations with Binet or similar test scales that are given individually to one. person at a time. When group tests, given chiefly in the schools, are added the number of yearly examinations increases hundreds of times. There is a score of different group tests by different, authors on the market. The sales on these run to several million copies a year. This hardly gives any foundation for the statement that "People do not believe in intelligence tests."

In order to get concrete information on this question, a questionnaire was sent out a year ago to various persons for whom we had been making mental test examinations. This included probate judges, county child welfare and federal relief agents, public school officials and teachers, agents of private social welfare organizations, and others, a total of 532 persons. The replies from 344 are tabulated and discussed in a brief article on "What others think of mental tests," In the Psychological Exchange of January 1935. The questionnaire asked each person to mark the one of five statements given which came closest to expressing his opinion about intelligence tests. The five statements were as follows:

These mental examinations arc: 1. Of but little or no value; 2. Of some value but not worth the costs; 3. Desirable and should be made; 4. Of much value and highly recommended; 5. Essential and necessary for my work."

The percentages of the 344 persons who marked each of the five statements were as follows: 1, 0.5; 2, 4.3; 3, 25.0; 4, 17.0; 5, 53.0.

This indicates that ninety-five per cent believed in mental tests to the extent, at least, that they thought they were "desirable and should be made." But more important is the fact that fifty-three per cent, a majority, felt that mental tests were "essential and necessary for their work."

The Nature and Construction of the Intelligence Test Scale: Let us consider the manner in which intelligence test scales arc devised. The methods used have become highly technical. Thousands of articles and scores of books have been written on the .subject. But it will be possible to give a few fundamentals in a simple, understandable way. If these are grasped it will be easy to compare the test method with other methods of determining grades of intelligence. Intelligence testing, as we understand it today, began with the Binet scale of tests published in 1908. Binet, with others, had been working on the problem for many years. All had attempted to solve it by analyzing intelligence into its supposedly separate parts and then attempting to measure those separate parts, such as capacity to make sensory discriminations, memory, association, attention, reasoning, and so on. Binet succeeded when he abandoned this procedure, and attacked the problem in an entirely different way. Giving up all definitions of intelligence, he set out to learn what children of different ages could do, to find tasks that six-year-old children could do but five-year-olds could not, tasks that seven-year-olds could do but six-year-olds could not, and so on up the age scale. He succeeded in finding such tasks sufficiently to construct a scale of tests with them, applicable to children from the ages of three to thirteen years. To determine the intelligence of a child or adult with this scale of tests all that was required was to give him these tests in order as far up as he could pass any of them. If, for example, he passed as many as the average seven-year-old child could pass he would have the mental development of the average seven-year-old child, for which Binet invented the term "mental age." Binet expressed grades of intelligence in terms of years of difference between age and mental age, those with mental ages less than their ages being below average intelligence, and those with mental ages greater than their ages being above average intelligence. We may leave Binet at this point. For Binet's chief Contribution was not the particular tests he devised, but, in the method of producing them, the use of age-norms, leading to the mental age concept. But before we pass on to later refinements, let us note carefully the simplicity of a Binet scale of tests, and its freedom from untenable assumptions. It uses no definition of intelligence and needs none. It does not pretend to measure any particular mental function or group of functions, and it is not necessary to know what is measured. It makes and needs only two assumptions. The first is that the mind develops during childhood. The second is that the increase in scores on the tests from one age to the next higher is due to this mental development chiefly or entirely, and not to something else, such as training. Everyone grants the first assumption. The second was not entirely correct for the original Binet scale, and much of the improvement in later tests lies in making it more correct. One might go further than this and contend that these two assumptions are needed only as a means of devising the tests. In the use and practical application of the test results neither of them is essential. The only thing we need to know about a test score is its meaning in terms of ability to do other things more directly concerned in everyday life. If we know, for example, that a certain test score for a child means ability to do third-grade work in school, and that a certain other score for an adult means ability to succeed in a certain occupation, we can make about the same practical use of the results as we could if we knew in addition that the test score in question represented a certain mental age and a certain combination of mental processes necessary to produce the test score. Among the most frequent questions asked are, "What does this and that test measure?" and "Of what value is the result if you don't know what it measures?" My answer is that 1 can usually make only a rough guess as to what a test measures, and that it would add very little to the usefulness of the result if 1 did know exactly.

The up-to-date present day intelligence test scale goes far beyond Binet in many important refinements, although nearly all are still based on its general principles. To produce a complete scale of tests for all ages from near birth to mental maturity that will have at least equal merits with others already on the market, is a big undertaking for any one individual. The labor involved requires an organization for a period of years to select and devise the tests, to secure norms for them at all ages, to complete the almost endless statistical work necessary to determine their merits and faults, and to finally round up the tests into a finished scale ready for practical use. This is why we have innumerable incomplete scales applicable at a few age levels only, and which are for the most part group tests that are much more easily devised, and only several relatively complete individual test scales that, are still up to date enough to be in use.

Preliminary Selection of Tests: The, first step in the construction of a test scale is the preliminary selection of tests to be tried out. In this selection three needs are kept in mind. First, the tests must be of various degrees of difficulty so as to fit the abilities of children at different ages. Second, the ability to perform the tasks set in the tests must be affected as little as possible by any special training or lack of it. They must measure mental development, and not something else that may vary independently of mental development. Third, the tests must represent as great a variety of mental tasks as possible. There must be an adequate number and none that duplicate others in what they measure. The source of the tests in this preliminary selection lies in all that psychology knows about mental development in children. In a large measure and more specifically it lies in innumerable tests that psychology has used for one purpose and another but which have not been incorporated into any complete scale of intelligence tests.

Establishment of Age Norms: The second step consists of trying out the tests selected to determine what scores children of different ages make on them. We call it the establishment of age norms. Aside from the labor involved in testing thousands of children, this seems simple. It is, however, not easy to get children for testing who are known to be truly representative of the general population of the whole nation with respect to intelligence. The average school child becomes increasingly above the average of the general population as we go through the school grades, because of the elimination of the dull through school failure. There are also local differences due to the predominance of some occupation or nationality of parents.

Final Selection. Validity: The third step consists of the final selection of the tests that are to be kept and combined into the finished scale. This involves a number of considerations. Minor ones are such matters as ease of administration, freedom from effect of repetition and coaching, and capacity to arouse and hold the child's interest. The two major considerations are the tests' validity and reliability. A large body of statistical technique has grown up to determine these two major aspects of mental tests. An intelligence test is fully valid when it measures intelligence and nothing else; that is, when scores on it rise and fall as intelligence rises and falls. We may say at once that there is no single and entirely satisfactory method of determining exactly how valid a test is. Statisticians as a rule have greatly over-simplified this problem as well as that of determining the reliability of a test. We determine validity by finding the amount of agreement between the scores on the test and some other criterion of intelligence. This is easy and would be satisfactory if we had any other criterion of intelligence that we know is more valid than the test scores themselves. School grades and marks, for example, are some criterion of intelligence. But psychologists soon learned that they were usually a poorer measure of intelligence than almost any set of intelligence tests are likely to be. Thus psychologists have accepted the rule that the scores on a good intelligence test must correlate, agree, only moderately well with school marks. Too high correlation would mean that the test scores were determined by the same things that determine school murks. And it is known that a child's school marks are determined by a number of things other than his intelligence.

In another method of determining validity we assume that the total or combined score on a number of tests always has considerable validity and more than the score on any one test alone. The last part of this assumption is certainly correct. With this we can determine the relative validity of the different single tests, and eliminate the poor ones that show a low correlation with the combined score on all the tests.

The amount of improvement in the test score from one age to the next gives still another way of determining validity. Since the purpose of the test is to measure the amount of mental development from one age to the next, it follows that a test is the better the more the score on it increases with increase in age, provided that it is not measuring something else that also increases with age. This something else is usually training, and the presence of training effect on test scores can be found in other ways. Improvement in the score on a test with increase in age gives, in my judgment, the best single method of determining validity.

Reliability: By reliability of a test is meant the consistency with which it gives the same result when it is given a second time to the same child. There are two methods commonly followed in determining reliability. One is to repeat the whole set of tests on the same children a second time shortly after they were given the first time, and then compute the amount of agreement, or correlation, between the scores from the two examinations. A similar procedure gives the tests only once, and then divides the group of tests given into two halves, and computes the correlation between the scores on the two halves of the group. This correlation may then be corrected to what it would most likely have been if the first method had been followed. The correct interpretation of these correlations is one of the most difficult tasks about intelligence tests. For there are circumstances under which these correlations will be high because the tests are poor instead of good. For instance, if the tests are too easy children will tend to get the same score a second time on them because nothing can interfere at either time to prevent their getting a maximum or near maximum score, which will then necessarily be the same both times. Among the things that may interfere and result in different test scores on the same children at different times are changes in attitude with resulting changes in effort the child makes, changes in emotional conditions that may handicap one time and not another, changes in health, changes in outside distractions during the testing, and others. If for the reason just given or any other a test score remains unaffected by these changes it is obviously a poor rather than a good test, yet it would show high reliability, as reliability is defined.

Units of Measurement: Passing over remaining details concerning the final selection of test that are to be incorporated in the finished scale, we meet the

next problem in determining the unit of measurement in terms of which lest results are to be expressed. The immediate test result is what we call the "raw score." This may consist of the number of tests passed in a group of tests, or of the number of items passed in a tingle test.. To add up these tests or items passed by a child for his total score, however, would be like adding up the length in inches, feet, and rods of different parts of a road to get the total length of the road. Raw scares cannot be used as units of measurement in finding the intelligence of children. But they can be converted into other units that are more accurate. Among these are five that have been extensively used or at least widely advocated. 1 shall at this point only enumerate and define them, with brief comments on their relative merits. Later I shall give some test results comparing the two in which we are most interested.

Binet converted the raw score into mental age. He graded intelligence by the number of years' difference between age and mental age. Thus the mental year of development was his unit of measurement. But it was recognized immediately that the amount of mental development during a year decreases very much as we go up in the age scale. Apparently the average child grows mentally four times as much from four to five as he does from fourteen to fifteen. Consequently four years of mental retardation at fifteen would be no more than one year of retardation at five. This resulted in the use of the ratio between mental age and age in place of Binet's differences between mental age and age; that is, the intelligence quotient, obtained by dividing mental age by age. The I. Q. has become the generally accepted unit of measurement in expressing grades of intelligence. It has two major faults, only one of which can be remedied, and I believe it will be discarded in the not distant future for a, better measure that is being proposed. Its first fault lies in the fact that mental ages cannot go higher than that of the average adult, so consequently I. Q.'s for adults cannot go above 1.00. This fault can be and has been roughly remedied by devices that compute I. Q.'s above] .00 for adults that are roughly equivalent to the I. *Q's* above 1.00 m children. The second fault in the I. Q. is its general tendency to decrease with age if it begins below 1.00, and increase with age if it begins above 1.00. The rate of change varied greatly with age and with how low or high it begins, so that corrections became too complicated to be practical. This change in I. Q. does not represent any change in intelligence, but is a fault in the I. Q. measure itself. It results from the fact that the rate of normal mental development decreases with age, making the mental year of development smaller and smaller as the child grows older. Obviously if we could find an accurate unit for measuring mental development in place of mental age we would have the means of getting an accurate quotient that would not change with age as does the I. Q. Such a unit has apparently been found by determining the exact nature of the normal mental growth curve. 1 attempted to find this growth curve from my data on the biannual reexaminations of inmates in the School for Feeble-Minded over a period of ten years, but failed to get one sufficiently accurate. [Kuhlmann, F. Results of Repeated Mental Re-examinations of 6,391 Feeble-Minded over a period of Ten Years. J. Appl. Psych., 1921] In 1926, five years later, Heinis, a Swiss psychologist, published a formula for a growth curve, and, using my data, showed that his quotients remained constant where the I. Q. changed with age. [Heinis, H. A. Personal Constant. J. Educat. Psychol. 1926] Heinis did not. give the values for his growth curve for each year and month, but only the formula from which they could be computed. In order to study the merits of this method 1 made these computations [Kuhlmann, F. A Median Mental Age Method of Weighting and Scaling Mental Tests. J. Appl. Psychol., 1927], which were later checked with some corrections and extensions by Hilden [Hilden, A. H.

Table of Heinis Personal Constant Values. Educat. Test Bureau, Minneapolis, 1933]. Giving them only for the years and not the months, these growth curve values are as follows:

Years	1	2	a	1	5	6	7	8		10	11	12	13	14	15	16	17	18	19	20
Values	60	111	155	193	226	254	279	300	318	333	346	358	368	376	3S4	390	395	400	404	407

Heinis' Quotient, which he calls the "Personal Constant," is obtained by first converting the mental age and age into the units of this growth curve, and then dividing the units representing the mental age by the units representing the age. For example, for a ten-year-old child with a mental age of five we divide the 226 by 333, getting a quotient of .67. The I. Q., of course, would be .50 in this ease. This quotient of .67 means that this ten-year-old child's mental development is .67 per cent of the average at the age of ten, when the units of measurement used are true and correct units instead of the mental age with its variable magnitude. I have renamed Heinis' "Personal Constant" the "Percent of Average," both because the latter is more easily understood and exactly describes the fact, and because the term "Personal Constant" is misleading in that it suggests that this constant remains the same for any and every particular individual, which is not the case at all. The true intelligence of any particular child may, within certain limits, vary up and down from year to year, and Heinis' quotient shows this variation correctly, where the I. Q. does not.

There are two other methods of expressing grades of intelligence that should be considered, not because they are frequently used, but because they are so constantly advocated. One of these is the percentile method. The number of tests or items in a test a child passes arc first added for a total raw score. Let us say this is done for each of a thousand six-year-old children. These scores are then arranged in order from lowest to highest. To get the percentile norms for these raw scores we then simply begin at one end, say the lower, and count off one, two, three, and so on, percents of the thousand scores, up to the end. Say ten per cent brings us to a raw score of 12 tests passed. This raw score of 12 is then converted into the ten percentile, and any other six-year-old child passing 12 tests and no more is given the ten percentile score. In the same way the raw score equivalents are found for every other percentile up to the 100 percentile. One fault with percentile scores lies in the inequality of the percentile intervals from one to one hundred. Children of average or near average intelligence are very numerous as compared with the number who are extremely dull or extremely bright. Consequently the true difference in intelligence between a five and a ten percentile child is many times as large as is the difference between a forty-five and a fifty percentile child. In fact, the percentile steps or intervals are quite too large at the extremes for practical purposes, and many times smaller than can be used at the fifty percentile level.

This fault has been corrected with the aid of the assumption that the frequency of the different grades of intelligence corresponds to the so-called "normal" distribution. The normal distribution curve is perfectly symmetrical, giving the highest frequency for persons of average intelligence with decreasing numbers belonging to grades below and above average. The base line for this distribution curve may-then be divided into a hundred or any other convenient number of equal steps or units. The number of cases falling under the curve at each of these steps is known, and this gives us a means of converting the raw scores on mental tests into a measuring scale with supposedly true, equal units. Thus, with the base line divided into 100 equal steps, and with the raw scores for our 1,000 six-year-old children arranged again in order from lowest to highest, we count off .02 per cent up from the lower end to get to the one percentile score. The ten percentile score brings us to .68 per cent of the 1,000 cases, and so on, giving the following scale:

Percentile	1	1	10	20	30	40	50	60	70	80	90	1OO
Per cant of total number	.02		.68	3.45	11.37	27 29	50	72 71	88.93	93 96	99.32	100

Instead of dividing the base line into 100 equal steps, other units are often used. Among these are Standard Deviation, the Quartile, and the Probable Error. It is unnecessary to explain these terms, as none involve any new principles, but only represent units of different magnitudes, as does the centimeter, inch, and foot. From a theoretical standpoint these measures of intelligence make a strong appeal. Granting underlying assumptions, they have none of the faults of the I. Q. It is probably for this reason that most textbooks recommend their use. But most of these textbook writers have never attempted to devise intelligence test scales, and have not met the practical difficulties these measures run into. These difficulties, common to these and percentile measures, are, in a word, as follows:

1. They give only a measure of intelligence, and none of the total amount of mental development up to date. They give us a substitute for the I. Q., but, none for mental age. For practical purposes, a measure of the total amount of mental development is as necessary as is a measure of grade of intelligence.

2. It is very difficult to get reliable age norms for this type of scale. Note that we must have a whole scale of age norms at each age, instead of only one for all ages. The raw score on the tests that represents the 50 percentile six-year-old does not represent the 50 percentile seven-year-old. Also note that it requires a relatively very large number of eases at any age to get reliable norms for the whole scale at that age. In scales measuring in terms of mental age and I. Q. or equivalents, a hundred eases, if carefully chosen, give a quite reliable average score or age norm. A hundred six-year olds would be entirely inadequate for a percentile scale. To attain any semblance of reliability at all points it would require at least a hundred times as many, or 10,000. Even this large number would give us the scores on only two children to determine the one percentile norm, and only 68 children to determine the ten percentile norm. Assume that these percentile norms are needed for about twenty different levels, one would have to examine 200,000 cases in order to establish all the norms needed for a complete intelligence test scale. That is four times as many examinations as the Division of Research has made in the twenty-five years of its existence.

3. A third difficulty with the percentile and similar measures of intelligence is with the necessary assumption that the distribution of grades of intelligence of children available for examination conforms to the normal distribution curve. This is not at all proven. In fact, it is hard to believe that this distribution is not seriously affected by the increasing eliminations of children from the schools as we pass up through the grades. If we assumed that only the dullest two per cent of the first grade were eliminated in the eighth grade, the 16 percentile child of the first grade would become the one percentile child in the eighth, when a percentile represents one one-hundredth of the base line of a normal distribution. If four per cent were eliminated, which is not an unlikely number, the 21 percentile child of the first grade would become the one percentile child of the eighth grade. When this matter is taken into account it is seen that the percentile is even less constant than is the I. Q. and no longer retains any advantages over it.

I have given this much time to discuss the nature and construction of the intelligence test scale in order to give you at least some rough impression of what is back of a simple brief statement that John Jones has been mentally tested and found to have an I. Q. of .70, and in order that we may have some basis for judging the relative reliability of this I. Q. .70 and the observation of the teacher, attorney, judge or other that, John Jones is quite normal. It has been claimed, and I think with good reason, that the intelligence test is the result of more work and painstaking effort to get an accurate and reliable measuring instrument than has been put into any other measure or instrument in any other field. This does not mean, of course, that we can therefore measure intelligence as well as we can space electrical forces, but it does mean that the intelligence test result is entitled to respect.

Other Methods of Measuring Intelligence: It would take entirely too long to give even the briefest outline of the other than test methods of measuring intelligence. The fact is that everyone who does not use mental tests has to a large degree his own procedure and his own ideas on what are criteria of intelligence. Furthermore, no one has any fixed standards to go by, such as the mental test age norm, and different examiners with the same observations before them come to entirely different conclusions as to the grade of intelligence they represent. Moreover, the psychologist has followed up about every conceivable criterion of intelligence ever used or suggested by the old-line experts and laymen as well, and his put them through the same rigorous scrutiny as has been given intelligence tests. They have been found too wanting in validity and reliability to be used alongside with the intelligence test. You can be assured that if the psychologist had found that family histories, health records, physical traits, school records, occupational histories, behavior records, impressions from personal interviews, and so on, gave trustworthy evidence and a satisfactory means of determining grades of intelligence he would have retained them, and no intelligence tests would probably ever have appeared. It is because the experts, be they teachers, physicians, psychologists, or what not, disagree so frequently and so widely on cases that wc have intelligence tests.

I said before that we determined the reliability of intelligence tests by giving them twice to the same children and then computing the amount of agreement between the results of the two examinations. It does not matter materially whether the tests are given both times by the s:tme or by two different examiners, if the examiners are reasonably qualified. We can in this way compare the reliability of the results of the expert observer with the reliability of the test results.

I shall close this subject with a few figures from one such comparison. One hundred fifty children in the school department, of the School for Feeble-Minded were selected for the teachers to grade as to intelligence. They were given the following instructions: Each teacher was to select her own list from the 150, ineluding only children she felt she could grade accurately. She was to spend two months to get further observations, and not discuss the matter with any other teacher. Most of the teachers had known the children they graded for a year or more. She was to class each child graded under one of five grades, which we may number 1, 2, 3, 4, 5. They were told that all children had been tested, and that the mental ages ranged from eight to twelve years inclusive. With perhaps a few exceptions they did not know what mental age any child had. After each teacher had selected her own list there were fifty cases left of the 150, each of whom was graded by at least three teachers. It is evident that, this gave an unusually favorable situation for reliable grading, far superior to what we usually have when the expert reads available records of a case, hears some testimony, and adds a personal interview with the case. It should be noted that the teacher of mentally defective children has better opportunities to become expert in judging and grading the intelligence

of these children than has any other person. The results may be tellingly expressed in a very few figures.

Range of differences	0	1	2	3	4
Number of children	7	6	19	9	9
Per cents................................ _	14._	12	38	18	18

The first line of figures gives the rank differences between the lowest and highest grade given a, child. The second line gives the, number of children under each, and the third line gives the corresponding percentages of the fifty cases. Thus for seven of the fifty children graded, all teachers agreed. For six children the teachers disagreed by one grade, as when the highest grade given was three, and the lowest two. Note that for eighteen percent the disagreement amounted to four grades, the equivalent of one teacher grading a child as a mental eight, while another teacher graded him a mental twelve. In terms of I. Q.'s this amounts to thirty points. iKuhlmann, F. The Binet-Simon Tests of Intelligence in grading Feeble-Minded Children. Journ. Psycho-Asthemr.s, 19121

No further comments are really necessary on these figures. With our later Binet type intelligence test scales a difference of thirty points in I. Q. on two different examinations occurs only in a very small fraction of one per cent of re-examinations When such a large difference does occur on re-examination of a case it is probably always due to "some major change in his intelligence or to some <u>outside</u> factor that has nothing to do with the reliability of the tests.

Some Results of Intelligence Tests: The main results of the intelligence tests made by the Division of Research are given in the biennial reports of the Board of Control. These need no further discussion here. You all know by this time that the delinquents in the reformatories and prison are on the whole below average intelligence, end that the same is true of children at the Owatonna State School and at the Gillette Hospital. Certain aspects of the results not given in the biennial reports are of importance in interpreting the I. Q.'s of individual cases. Let us turn to these next.

Tendency of I. Q. to Change and the Heinis Measure: It was noted above that the I. Q. has a general tendency to change, if it begins below or above 1 00 and that another measure of intelligence developed by Heinis corrected this error More details on this will have a highly practical value in putting the I. Q. to use in dealing with the individual child. My re-examinations at the School for Feeble-Minded involved 639 cases. These showed marked decreases in the I. Q. during the ten-year period. For the same data Heinis' scores, which I shall call the per cent of average, showed no important change. The following table gives the figures making the comparison. [Kuhlmann, F. What the I. Q. menus today. The Nations' Schools, Feb., 1933. The figures given here include slight correction of those given by Heinis, due to his error in assuming that the ages in my data were 9.5, 10.5, etc., instead of just 9, 10, etc. They are also extended to ages 7 and 8. by smoothing out my original figures]

	AGE	7	8	9	10	11	12	13	14	15	16
Borderline	I. Q.	83	81	79	78	76	75	73	71	69	66
	Pc. Av.	89	89	88	89	89	89	89	89	88	88
Moron	I. Q.	66	64	63	63	62	60	59	57	55	53
	Pc. Av.	77	77	77	78	78	79	80	80	79	80
Imbecile	I. Q.	42	41	40	39	37	36	34	33	32	30
	Pc. Av.	55	55	56	57	57	57	57	58	57	56

You will note in these figures that the I. Q. changes most for borderline cases and least for imbeciles, while the per cent of average varies a maximum of three points only, with no tendency to a continual change in one direction, where the I. Q. decreases a total of seventeen points from age seven to sixteen. While we have as yet no figures for younger children, we also have no reason for believing that they would be very different. If there is no material difference for younger children it is obvious that the I. Q. for a young child may give a very misleading idea of what his intelligence will be at maturity. Apparently children of pre-school ages with intelligence quotients from .80 to .90 would tend to become true morons by the age of sixteen.

Hilden obtained quite the same results on this comparison, using the accumulated re-examinations made by the Division of Research during the second ten-year period. [Hilden, A. H. A Comparative Study of the intelligence quotient and Heinis' Personal Constant. Journ. Appl. Psychol., 1933. Also, Table of Heinis' Personal Constant Values. Educat. Test Bureau, Minneapolis] The following table is taken from one of his graphs.

Years Interval		0	1	2	3		4	5
Change	I. Q	0	1.7	4.4	1.4	6.2	7.8	7.8
	Pc.. AV.	0	0	—1.	—0.3	—0.5	0	+0.5

The figures are averages for a group of cases of various grades of intelligence. The I. Q. decreases about 1.5 points a year, while the per cent of average again shows no tendency to change in any direction, and has a maximum variation of Only half a point. Several other studies, which I shall not take the time to quote, give substantially the same result on the superiority over the I. Q. of the Heinis measure.

Another difficulty with the I. Q. concerns chiefly the I, Q.'s of adults who are above average intelligence, but is met also with children of very high intelligence. Let us assume two twelve-year-old children, one fifty per cent below average and the other fifty per cent above average, when these per cents are based on true units of measurement. If they were that, it is obvious that the mental age of the bright child would be much more above twelve than the mental age of the dull child would be below twelve. The mental ages would not be six and eighteen, but less than six for the dull and much more than eighteen for the bright one. According to the Heinis mental growth curve a true fifty per cent below the twelve-year level gives a mental age of three yrs., seven mos., while a true fifty per cent above the twelve-year level would give a mental age of several hundred. The point I wish to make is that the very high I. Q.'s obtained on adults and for older children are misleadingly high. At that they are much lower than they would be if we had tests that scored high enough correct mental ages as such. If we had such tests we would get mental ages and I. Q.'s of entirely ridiculous and meaningless magnitudes. The new tests that we began to use a year and a half ago give more discriminating scores than the old at the extreme upper mental levels of adults. When we convert the raw scores into I. Q.'s at these points we sometimes get I. Q.'s of over 3.00. For such cases a percentile score would probably be less misleading. But I have hopes that the, Heinis score in terms of per cent of average can be developed to a point where it will overcome the fault? of both I. Q. and percentile. In the meantime please do not take reported I. Q.'s of over 2.00 too seriously.

The occasional large difference in I. Q. that we get on the re-examination of a case has worried both psychologist and those who want to use the test result perhaps more than anything else. There is less occasion to remember the frequent agreements, and, as usual, the quite infrequent large exceptions produce the general impression that they are the rule. There are a great many things that cause a change in I. Q. on reexamination besides the fault in the I. Q. as a measure and the unreliability of the- test score from which the I. Q. is derived. Let us look first at some figures on re-examination changes just as found. A consideration of the factors that produce changes will then help to interpret them, and give us, I believe, a pretty concise idea on how much change we may contribute to unreliability of the tests and how much to other factors. The next table gives re-examination results on all the re-examined cases who were not less than live years old.

I. Q. changes	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
0-2 yrs	64	116	92	98	56	65	57	43	35	26	15	11	5	5	7	2	1	1	3	1	2		1	3
3 yrs- and over	35	71	71	68	70	49	51	51	55	33	23	24	25	24	15	13	15	9	14	11	5	5	l	5

I. Q. changes	24	25	26	27	28	29	30	31	No.
0-2 yrs				1	1				721
3 yrs. and over	3	1	1		3	2	..	1	764

The first line of figures gives the number of points change in 1. Q. on the re examination. The second line gives the number of cases when the interval be tween one examination and the next was from 0 to 2 yrs. inclusive. The third line of figures gives the number of cases when this interval was 3 yrs. or more Thus when the interval was 0-2 yrs., 64 out of 721 re-examinations gave exactly the same I. Q. as did the previous examination. There was one point change i I. Q. in 116 cases, and so on. When these figures for the 0-2 yrs. intervals are converted into percentages, it seems fair to deduce the following general rule on expected changes for these intervals:

25 per cent of examinations will give not over 1 point change; 50, not over 3 points change; 75, not over 5 points change; 96, not over 10 points change-

In these figures I have not included re-examination results for children be than five years old because the tests used for these younger children have been changed quite a number of times and the results are not comparable. On the whole, however, it seems that for the same interval between examinations the changes in I. Q. have been about twice as large as for older children. This brings us to the question of the different factors that cause these changes. So far have mentioned only the fault, in the I. Q. as a measure and the unreliability the tests.

Probably the biggest single factor causing changes is the difference in effort made by the person examined during the two examinations. This is still one thing that the tests and the examiner can control only rather imperfectly. On the whole, the younger the child the harder it is to get his best effort in an examination. On the whole, also, the younger the child the more his effort will depend on the inherent interest the tests themselves can arouse. Thus the mental test for the child of pre-school age is made mostly through the medium of toys and games carefully selected and adjusted to the predominant interests at each particular age. With older children and adults we rely more on the examiner's skill to give the examinee a special motive for making his best possible record. With very young children, interest in toys and game3 depends to a high degree on passing moods and disposition over which the examiner has no control. Could the examiner pick his time and circumstances for examining either child or adults, changes in results on re-examination might perhaps be reduced fifty per cent or more. The psychologist may prescribe conditions required for an examination, but in practice our examiners are compelled to violate all of them repeatedly. They must examine babies when they are tired from a long ride to the place of examination, or are sleepy or hungry because a nap or feeding *is* overdue, or when made fretful by an unskilled attendant, or are suffering some physical discomfort, or are disinclined through minor nutritional and health condition, and so on. They must examine older children and adults frequently when they are at the peak of emotional upsets, as when going to court for some misbehavior, or are threatened by some dire change in their lives, such as commitment to a state institution, separation from their parents or friends, entrance into special class for mental defectives, and similar calamities. When they have been in conflict with parents, teachers, employers, social workers, and courts and have become violently antagonistic to everything and everybody the mental examiner is called in to "examine their heads" to see what is the matter with them. To get their cooperation for a reliable test result gives the examiner a real job.

Another cause of changes in I. Q. on re-examination is actual change in intelligence which the tests then correctly measure. There was once a strong conviction that there never was any recovery from feeble-mindedness. Mental tests have shown that recovery can and does sometimes take place, as well a marked deteriorations. Among cases that have had repeated re-examinations we find a few of both classes, with the deteriorations much more frequent than the improvements. The following cases are given merely as illustrations. [Taken from "What the I. Q. means today," cited above] We have no adequate data to indicate how frequently they occur.

groups to which we then give different treatment because of this classification, such as the feeble-minded and normal. Within either of these groups, especially the normal, we pay little or no attention to a difference of ten to fifteen I. Q. points.

We come now to the last question to be discussed. Even assuming that the intelligence tests always give exactly correct results on intelligence, the children and adults that come to state institutions as a rule do not behave in accordance with their intelligence. If they had done so, most of them would not have been sent to an institution. This comes near to being true even of the feeble-minded. That intelligence is not the whole personality we knew long before we had intelligence tests. Now when we have I. Q.'s we sometimes forget this, and question the accuracy of the I. Q. when it does not explain behavior. And vice versa, we tend to call people feeble-minded or normal according to their behavior, rather than according to their intelligence. Under present laws and with existing institutions this is often not only defensible but the most sensible procedure to follow. Certainly there are plenty of cases with I. Q.'s in the eighties who are as damaging to society as any others with much less intelligence. The fact that is of immediate practical importance is their incompetence and delinquency. It matters little whether this is due to low I. Q. or to just bad habit, so long as the bad habit is a relatively fixed part of their personality. "Under existing conditions it would be more sensible to commit such cases to the institution for feeble-minded, and then discharge them as normal if and when their bad-habits or other personality traits responsible for their behavior were sufficiently improved. For a similar reason it might be more sensible to do nothing about adults with I. Q.'s in the sixties so long as they remained well behaved and needed no public support. But one can hardly subscribe to this failure to discriminate between the different causes of incompetency and delinquency as a permanent state policy. It could only lead to the limiting of treatment to custodid care, and to the elimination of attempts to learn and understand the total personality of the individual with the differentiation of treatment necessary to improve and reform. And in the meantime it would soon make chaos of our institutions even as wo have them. For I. Q.'s of sixties and of eighties do not mix even in our present institutions for the feeble-minded. The establishment of a more specialized institution for the defective delinquent is, of course, at least a partial solution to this problem. And we will never solve it this way if we proceed to commit persons of relatively high I. Q.'s as feeble-minded simply because their behavior has become intolerable.

The question of lack of agreement between I. Q. and behavior or achievement comes up in another form with cases after commitment to the institution. Children of the same mental ages in the institution's school show a wide range in the quality of school work they do, and those with the higher mental ages often do more poorly than others with lower mental ages. This is because some of the children in institutions have had more handicaps than children outside. Many have physical and health handicaps that do not affect mental age as much as they do school work. Others with perhaps relatively high mental ages are poor in school because of irregular and poor attendance in school before they come to the institution. Still others do not do as well as might be expected of their mental ages because of poor motivation and lack of interest acquired in the poor homes from which they come.

Still other and more factors appear to disturb the relationship between intelligence test results and the work of the adult institution inmate in the various occupations within the institution. With the adult, bad personality traits have had more time to develop. Bad habits have increased and become fixed. Social contacts resulting in chronic antagonisms have been made. Desires and ambitions detrimental to good behavior and achievement have appeared. Many feeble-minded as well as delinquents come to the institution with more or less fixed hatreds against the institution and officials that restrain them. The institution inmate can hardly be expected to live up to his intelligence as well as do others outside. For that matter, let us not forget that no one's conduct is always fully guided by the intelligence that he has. Everybody acts like a moron at times, and often all the time in some particular. Likewise the feeble-minded and the delinquent are usually not feeble-minded or delinquent in everything. But there are special difficulties in determining how well the institution adult does in a given occupation in the institution. In the first place the record of the quality of his work is made usually by some untrained attendant, who is often easily prejudiced and always unskilled in making accurate observations. In the second place the mental age or I. Q. needed for satisfactory performance in a given occupation is usually not known. And higher intelligence than is needed for the job will add but little to the performance. A man with an I. Q. of .60 may do as well with a shovel as another with an I. Q. of 1.20. Performance will not agree with 1. Q. We need to know much more than we do about what intelligence is needed in different tasks before we can judge accurately how much the 1. Q.'s agree or disagree with performance. Finally, when achievement or conduct and I. Q. disagree there is all the more need of knowing the I. Q., for then only is there a possible opportunity for improvement.

Mr. Foley: Thank you, Doctor Kuhlmann. Your paper was an exceptionally fine one. We appreciate the time and thought and study which you must have put upon it, and I am sure most of us will want to study it carefully when it appears in the Quarterly.

I want to thank you all for being present today and for staying throughout the entire program. It has been a long one but a very successful one, I think.

Let us adjourn, until we meet again.