DOCUMENT RESUME

ED 048 330 TM 000 382

AUTHOR	Humphreys, Lloyd G.; And Others
TITLE	Project on Techniques of Objective Factor Analysis.
INSTITUTION	Illinois Univ., Champaign.
SPONS AGENCY	Office of Naval Research, Washington, D.C. Psychological Sciences Div.
PUB DATE	Apr 70
NOTE	117p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$6.58
DESCRIPTORS	Ability, Correlation, *Factor Analysis, *Factor Structure, Group Intelligence Tests, Group Tests, Homogeneous Grouping, Individual Tests, *Intelligence, *Personnel, Predictive Validity, *Psychological Tests, Psychometrics, Test Construction, Test Reliability, Test Validity

ABSTRACT

This collection of papers, concerned with the nature and theory of intelligence, forms part of a project to integrate test and factor theory with the empirical, functional relationships involving standard intelligence tests. The project will render more objective the use of factor analysis in personnel research. A definition of intelligence encompassing biological and socio-psychological factors is formulated in "Theory of Intelligence." Three classes of hypothesis are presented in "Hypothesis Developed From the Theory." In "The Psychological Test" a psychological theory is delineated as a basis for developing a theory of intelligence congruent with the experimental and observational correlates of measures of intelligence. Interrelations of homogeneity, reliability, and validity are considered in the paper of this name. "Illustrations of Test Characteristics by Means of Physical Analogues," "Evaluating the Importance of Factors in any Given Order of Factoring," and a description of "The Scottish Survey of Intelligence" complete the collection. (Author/CK)

TR 1 OCT 19 1970

PROJECT ON TECHNIQUES OF OBJECTIVE FACTOR ANALYSIS

The several separate manuscripts accompanying this preface were destined to be chapters in a monograph on the nature of intelligence. The plan was to integrate both test and factor theory with the empirical, functional relationships involving standard tests of intelligence. The completion of this work has been delayed by a sudden shift in career of the author.

This work was supported in part by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under Contract N00014-67-A-0305-0012 and in part by a sabbatical year supported by the University of Illinois. It is being forwarded now as a technical report under the contract in the hope that, even in its incomplete state, the material contained therein will be of use to personnel research activities.

Lloyd G. Humphreys
National Science Foundation

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Dr. Lloyd G. Humphreys, Department of Psychology
University of Illinois, Champaign, Illinois 61820

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
Project on Techniques of Objective Factor Analysis

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical Report

5. AUTHOR(S) (First name, middle initial, last name)
Lloyd G. Humphreys

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
September 9, 1970	100	38

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-67-A-0305-0012 b. PROJECT NO RR 006-04 c. Task Area No. RR006-04-01 d. Work Unit No. NR 150-305	TR 1 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research Department of the Navy Arlington, Virginia 22217

13. ABSTRACT

The initial thrust of the work under this contract was toward the development of empirically tested procedures designed to make more objective the use of factor analysis in personnel research. Later work has moved into more substantive applications. Ability theory, for example, has been closely tied to factor analytic techniques from the earliest work of Spearman in this field. What in factor analysis can and cannot do, and the confidence that one can place/the results are tied intimately to modern ability theory.

FORM NOV 65 1473 (PAGE 1)
S/N 0101-807-6801

UNCLASSIFIED
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Intelligence, Theory of Intelligence, Hypotheses from Theory of Intelligence, Psychological Test, Homogeneity, Reliability, Validity, Test Characteristics, Physical Analogues, Factors, Order of Factoring.						

THEORY OF INTELLIGENCE

It is necessary, as a first step, to formulate a definition of intelligence. The usual criterion for a definition is, of course, that the term in question in conjunction with other terms in the theory lead to testable hypotheses. The definition must lead to scientifically useful consequences. It is also reasonable to employ a secondary criterion on occasion. Since intelligence tests are in common use, and since these tests have become firmly entrenched in this society, the definition of intelligence should be tied directly to available measuring devices. This second criterion is compatible with a philosophy of science that does not dictate an operational definition for every concept in the theory, but it is more convenient to have operational definitions for certain terms in the theory than for others.

Definition of Intelligence. Intelligence is defined as the entire repertoire of acquired skills, knowledge, learning sets, and generalization tendencies considered intellectual in nature that are available at any one period of time. An intelligence test contains items that sample the totality of such acquisitions. Intelligence so defined is not an entity such as Spearman's "mental energy." Instead the definition suggests the Thomson "multiple bonds" approach. Nevertheless for the sake of convenience intelligence will be discussed as if it were a unitary disposition to solve intellectual problems.

There is one important difference from Thomson's multiple bonds, at least as the Thomson theory has at times been interpreted, that should be clarified. It is not essential that the person whose intelligence is measured have acquired a specific response to each stimulus or set of stimuli presented. Learning sets and generalization tendencies were introduced in the definition to preclude critical interpretations of this type.

The definition of intelligence here proposed would be circular as a function

of the use of "intellectual" if it were not for the fact that there is a consensus among psychologists as to the kinds of behaviors that are labelled intellectual. Thus the Stanford-Binet and the Wechsler tests can be considered examples of this consensus and define the consensus. It is also true that a present consensus does not rigidly define intellectual for all time to come. One should expect change to occur. This change will come slowly, however, because the process of changing the definition of a test in terms of the items composing it is a slow one. As the empirical basis for change primary reliance must be placed on functional relationships involving the total score on the test.

Contrast with Older Operationalism. This definition differs from the statement that intelligence is what intelligence tests measure. When the intercorrelations of several different intelligence tests do not approximate unity closely after correction for attenuation, the strict operationalist is left with as many different definitions of intelligence as there are tests. From the present point of view, however, one would not expect different tests to be perfectly correlated since each samples a domain that is fairly heterogeneous with a limited number of items. Parallel forms of the same test should be more highly correlated than different intelligence tests since in the former there is no item sampling error and there is near identity of parallel items.

A problem arises in trying to set a desired height of intercorrelations of tests sampling from the same domain. There is no easy answer. An a priori approach is not possible since a great deal depends on the number of items in each test and the degree of homogeniety of the domain. A combination of a rational analysis of the content of the tests in question plus a distribution of the intercorrelations of the proposed tests provides a partial answer. Tests of satisfactory reliability but whose correlations with other intelligence tests are not a part of the main distribution of such correlations can be considered

inadequate representatives of the domain. By this criterion a typical culture fair test of intelligence is not an acceptable measure of intelligence at this point in time.

A second difference between the two approaches to definition is that the present one fits into a larger context. Knowledge of learning and of the constitutional bases for learning become important. As a result the definition here proposed leads to testable hypotheses concerning intelligence.

A third difference between the present definition and the older, more superficial operationalism is that a distinction is made between the repertoire of responses, which is intelligence as here defined, and the eliciting of those responses on the test. A person whose repertoire of responses is for some reason not available at the time the test is administered can still be intelligent. This distinction is often phrased in the psychological literature as that between learning and performance, but the emphasis here is between acquired knowledge and skill, on the one hand, and performance on the other.

Discrepancies between intelligence and performance on an intelligence test can conceivably arise in a very large number of ways. The test constructer and the test administrator try to minimize the discrepancies by writing reliable, unambiguous items, by standardizing the conditions of test administration, and by specifying the populations of persons and the set of situations for which the test is appropriate. How successful such efforts are is an empirical matter and cannot be evaluated in the arm chair. A useful generalization from a great deal of such research is that intellectual performance is relatively robust. It is not affected substantially by many of the a priori possibilities. This finding should not, however, be taken as an excuse for careless or unsophisticated use of intelligence tests.

Biological Substrate. Since most theorists have defined intelligence as a capacity, generally fixed by inheritance, it is necessary to specify the reasons why this seems undesirable. It should be clearly understood at the outset that the present writer does not exclude the possibility, or rather probability, that constitutional differences among men affect the ease with which intellectual dispositions are acquired. He prefers the term "biological substrate" for intelligence to cover these differences while intelligence is reserved for the acquired disposition.

Biological differences can arise from many causes. In addition to genetically determined differences, biological differences can be acquired prenatally, perinatally, and postnatally. Furthermore, the genetically determined differences are far from unitary. Instead the genes are responsible for a huge complex of anatomical and biochemical factors. It is extremely doubtful that physiological psychologists are going to find a single key to the differential facility the human possesses in the acquisition of intellectual dispositions. Biological substrate and genetic substrate, respectively, for intellectual performances are more appropriate terms than a word which suggests an entity.

From the point of view of the user of an intelligence test the most important reason for not defining intelligence in terms of a genetic substrate is that a given person's standing with respect to genetic factors can not be inferred from a test score. The test measures acquired behavior. Independent assessment of the genetically determined biological base is presently possible for only a tiny portion of the human population, e. g., phenylketonouria. Some few of the acquired organic differences can be independently assessed, e. g., certain of the birth "injuries." Experimental control is lacking in studies of human genetics so that it is even impossible to draw conclusions about relative contribution to variance

of genetic factors in an snalyaia of variance design.

The conatruct of a genetic substrate for intelligence is required more by general biological knowledge and belief in biological continuity from lower animals to man than by good information concerning human genetics. Family relationship and other experimentally uncontrolled studies of human genetics are suggestive but not convincing. It is difficult to believe, however, that the controlled breeding studies of behavioral traits in lower animals could not be duplicated with the human if controls were posaible. More basic to this line of reasoning is the inference that any inter-species difference will also show intra-apecies differences. There are clearcut differences between man and other primates in the genetic substrate for intelligence. It is reasonable to assume that individual men will also differ in their genetic substrate for use of aymbols, abatract reasoning and problem solving, etc.

While a biological aubstrate for intelligence is made necessary by biological knowledge, the construct can not at the present time enter into testable hypotheses in any except the most general fashion. Any given organism may have innate capacity for the development of his intelligence, but the limits of this are very nebulous indeed. This capacity, furthermore, is not necessarily fixed at a given level throughout the life span. There may be genetically determined differences in the rate of maturation and of decline of the biological aubstrate that will influence individual differences in intelligence. It is safe to conclude that no amount of training will transform a chimpanzee into a human being intellectually, or a Mongoloid into a genius, but present data do not allow much more specific inferences than these.

Psycho-social Substrate. For basically the same reason that a test user can not draw inferences concerning genetic causes from a test score, he can not draw

inferences concerning environmental causes from a test score. Each human being is biologically unique. Two different biological organisms developing in seemingly identical environments will acquire different intellectual repertoires. Identical biological organisms developing in different environments will also acquire different intellectual repertoires. It is also true that similar repertoires can result from different mixes of heredity and environment. It is useful, therefore, to define a concept parallel to the biological substrate: namely, the psycho-social substrate. The psycho-social substrate for intelligence is just as important as the biological substrate, but is almost equally difficult to assess independently. Furthermore, the two are by no means orthogonal. Probable genetic differences among social classes, for example, accompany psycho-social differences.

It was stated earlier with respect to the biological substrate that only the most general sorts of inferences could be drawn legitimately. The same is true concerning the psycho-social substrate. If a man were raised in isolation, his intelligence would be very low. Quasi experimental approaches to this condition are furnished by canal boat and gypsy children (Anastasi, 1958). It is also probable that one could increase the quality of the psycho-social substrate with respect to developing intelligence and obtain an increase in intellectual level, but relatively little is known experimentally about this matter. Again, a quasi experimental approach to this problem is furnished by the comparison of intelligence of World War I and World War II draftees (Tuddenham, 1948) and of the World War II and 1963 norms of the Air Force Classification tests (Tupes and Shaycoft, 1964). The results are quite dramatic. Between the two World Wars, the increase amounted to approximately one standard deviation of the World War I distribution while subsequent to World War II the increase appears to be about one-half of a standard deviation.

In summary, response acquisition requires both a biological (including genetic) substrate and a psycho-social substrate which interact throughout the life span. Responses are acquired, and lost, during development, maturity, and decay. The test user can not draw specific inferences from a subject's test score about either of the two substrates.

Types of Behavioral Repertoires. A distinction is drawn traditionally between intelligence and achievement tests. A naive statement of the difference is that the intelligence test measures capacity to learn and the achievement test measures what has been learned. But items in all psychological and educational tests measure acquired behavior. The measures of even the simplest sensory and motor functions require a background of learning in order for the examinee to understand the directions and to provide answers.

A statement that recognizes the incongruity of a behavioral measure as a measure of capacity is that intelligence tests contain items that all examinees have had an equal opportunity to learn. This statement can be dismissed as false on its face. The psycho-social substrate is simply not equal for all. Opportunity depends on the characteristics of father and mother, siblings, other relatives, friends, the neighborhood, the schools, and other environment. There is no merit in maintaining a fiction. There is also no merit in belaboring this fiction as an argument against the use of tests.

Intelligence is here defined as the totality of responses available to the organism at any one period of time for the solution of intellectual problems. Intellectual is defined by a consensus among psychologists. The intelligence test samples the responses in the subject's repertoire at the time of testing. So defined, there are no differences in kind between intelligence and achievement, or between aptitude and achievement. There are instead three dimensions appro-

priate to the description of tests and the repertoires they sample (Humphreys, 1962). There are quantitative differences among different types of tests on these dimensions.

1. The most important of these dimensions is breadth. An intelligence test is much broader in coverage than individual achievement tests. Concurrent correlations between intelligence and achievement in a specific subject matter are quite high, but far from perfect. When a number of achievement tests in different subject matters are administered, thus achieving greater breadth on the achievement side, the total score obtained from the test battery is very highly correlated with measured intelligence. As a matter of fact, this correlation is about as high as the intercorrelations among recognized tests of intelligence.

2. A second dimension of difference is the extent to which a test is defined by a specific educational program. The achievement test is tied to a particular academic curriculum while the intelligence test samples both learning in school and out of school. An achievement test must be revised when the course of study changes while an intelligence test is more independent of what is being taught in a particular school at a particular period of time. The psycho-social substrate for the achievement test is more narrowly defined.

3. A third dimension of difference is the recency of the learning sampled. The achievement test measures recent learning primarily while the intelligence test samples older learning. Thus 8th grade arithmetic is a part of the "aptitude" section of the College Board tests and high school algebra is tapped by the "aptitude" section of the Graduate Record Examination, but similar questions administered in the 8th or 9th grade would be achievement items.

The use of aptitude requires additional clarification. The term is used commonly for one of the components of general intelligence as well as for an [illegible] not considered a component of intelligence. The former is the sense of

its use by the College Board and the Graduate Record Examination. Aptitude is also used at times in a very general sense to include both intelligence and non-intellectual abilities. No matter how used, however, there is no problem in fitting aptitude into the present analysis of differences among test items and the behavioral repertoires they sample. When used narrowly, aptitude and intelligence tests differ on the first dimension, but not on the second and third. Both aptitude and achievement tests would be classified as narrow, but an aptitude test in contrast to an achievement test assesses older learning that is not restricted to the classroom.

The dimensional analysis is useful in indicating why there is confusion concerning the proper category in which to place certain tests. Just because differences among test items are quantitative and not qualitative, it is possible for one man's intelligence test to be another man's achievement test. Thus Jensen (1968) categorizes the National Merit Scholarship Examination as an intelligence test, but precisely the same items are used in the Iowa Tests of Educational Development for assessing achievement. Frequently, the distinction between achievement and intelligence (or aptitude) tests is stated in terms of the purpose for which the test is used (Wesman, 1968). Purpose is independent of type of item. A test used for the prediction of future performance is called an aptitude test while the same test used to evaluate learning is called an achievement test. Thus, there is no conflict between the present definition of intelligence and the types of items used in measuring achievement and aptitude.

Contributions of Learning to Theory. Several different well established principles of learning contribute to the theory of intelligence being developed. The principles that are most useful are very broad and are also independent of the nuances of various learning theories. They might be said to be within the public

domain of accepted psychological knowledge.

1. One of the most important principles of learning for the development of intelligence is the presence of an intellectual psycho-social substrate. No one can learn to use abstract words who has had no contact with language. In the school the parallel principle is that of curriculum. A student will not acquire mathematical knowledge and skills who has had no exposure to mathematics. Note, furthermore, that it is exposure, not adequacy of exposure, that is the issue. In experimental attacks on type of exposure, type makes little contribution to variance. There are many cases also in which the exposure was highly ideosyncratic, e. g., Abraham Lincoln studying by fire light.

2. There must be motivation or incentive to learn. Motivation may be positive or negative, intrinsic or extrinsic, but must be present in some form. This statement of principle is intended to avoid an issue important in the psychology of learning. While reinforcement for some theorists is an essential part of the mechanism of learning, for others reinforcement is necessary for performance but not for learning per se. Nevertheless, all theorists acknowledge the importance of motivation for increased effectiveness of performance. Latent or incidental learning may exist, but it is very inefficient, and motivation is required for performance.

Given the fact that children differ in the type and degree of motivation for intellectual learning at a given moment in time, what is the source for these differences? There are again biological and psycho-social substrates for motivation as well as for intelligence. In this case the psycho-social substrate includes both the reinforcement history and current situational factors. In the absence of ability to manipulate the genetic substrate, for one who is interested in changing the course of future learning the necessary procedure is to control

type of exposure and to reinforce the behavior desired.

3. Forgetting is very slow for well learned or overlearned behavior. Given occasional rehearsal of learned behavior, practically no forgetting occurs. This means with respect to the development of intelligence that the intellectual repertoire continues to grow as long as the subject remains in an intellectual environment. This environment does not need to be an academic environment since an educated man cast away on an uninhabited island with a set of encyclopedias could still remain in an intellectual environment. There will be so little loss, in comparison with gain, for students during the school years that loss can be disregarded. For purposes of assessing the gain a total score uncorrected for differences in chronological age must be used; i. e., mental age units are adequate, but intelligence quotient units are not. With respect to the latter a person who does not show as much growth as his fellows will show a loss in I. Q.

4. Transfer of training takes place typically within a domain that the man on the street would consider quite narrow. In general measured transfer turns out to be less than nonpsychologists assume will be the case. For the development of intelligence this means that a great many relatively specific learnings have to take place. Primates can develop learning sets, but Harlow's monkeys learn relatively narrow sets (Harlow, 1949), e. g., the odd stimulus among a set of three. It takes each monkey a relatively large number of trials to acquire each such set. While the human brings to the learning situation a different and more efficient constitutional substrate for the acquisition of learning sets, or concepts, than does the monkey, it is still necessary for the human to acquire a very large number of these within the intellectual domain. (The number of these in the human is indicated roughly by the size of his comprehensive vocabulary.) While he does not have to acquire separately and individually each specific response

that psychologists would label intellectual, even the number of learning sets or generalization tendencies is very large so that a great deal of time is required for the learning.

5. Transfer is not only fairly narrow, but it can also be both positive and negative. Proactive inhibition is just as important as proactive facilitation. Or, to revert to terms that are more common in the literature of individual differences, a person can as readily acquire a disability as an ability. Certain disabilities are quite stable and quite resistant to change. Thus every person acquires to a greater or less degree a disability to speak a foreign language without accent. Few adults are able to overcome this disability. There are a very large number of items in the intellectual repertoire and each of these has both positive and negative effects on future response acquisition.

Contribution of Biology to Theory. Again, only the most general principles will be described. Unfortunately, the number of principles and their specificity in this area are not as directly pertinent to the development of intelligence as are the principles of learning. This arises because of the difficulties attendant upon doing controlled experimental work on the functioning of the human central nervous system and upon human genetics.

1. The companion principle to the first learning principle is that the subject must have a minimally adequate biological substrate. Persons showing the lowest levels of intelligence typically have biologically inadequate organisms. Children with phenylketonouria, Mongolism, cretinism, etc. will not be able to acquire intellectual behavior at a normal rate. Their capacity to learn is not well defined, and can be drastically underestimated, but capacity is none the less limited by their biological limitations.

2. The important distinction between phenotype and genotype is meaningless unless there is independent assessment of the genotype. A diagnosis of genetically

determined feeble mindedness from a test score is not possible. The combination of psycho-social and biological substrates leading to performance at the moron level may differ widely from one person to another who test at that level. It is useful at this point to repeat the injunction presented earlier: namely, it is impossible to draw causal implications concerning any substrate from the test score alone.

3. Each human being is biologically unique as a function of the number of chromosomes and number of genes in the genetic substrate and the large number of biological effects of events in the prenatal, perinatal, and postnatal environments. It is not even necessary to exclude monozygotic twins in making this statement, although the uniqueness of genotypes must be discarded for such twins. In spite of the uniqueness of genotypes, it is also true that there is a clustering of sorts among genotypes. This arises from the partial segregation of gene pools in sub-populations of the human species.

4. The biological substrate for intelligence includes a very large number of specific anatomical structures, physiological functions, and biochemical agents. It is highly probable that there are genetically determined individual differences in each of these and that these individual differences are for the most part independent of each other. The characteristics of all synapses in a given organism can probably not be determined from those of a particular synapse, or the characteristics of one ganglion in a given organism are not those of all ganglia. There are also possible a multitude of environmental effects on the biological organism that start at the moment of conception and extend throughout the life span.

Developmental Principles. There are at least two important principles for a theory of intelligence that can not be clearly distinguished as either learning

principles or biological principles. Both maturation and learning are presumably involved.

1. A person's present behavioral repertoire is an imperfect predictor of a future repertoire. This principle has been well documented by Fleishman and associates for motor learning (1954, 1955, 1960). Early trials are not correlated nearly as highly with later trials as adjacent trials are to each other. For the intellectual repertoire the principle has been substantiated by Anderson (1940) and Roff (1941). These latter investigators found that gains in mental age from year to year were independent of the base mental age at the start of the year.

There is ample a priori rationale for this principle. There is a great deal of seeming randomness in anyone's environment which will affect the psycho-social substrate and even at times the biological substrate for intelligence. The school a child attends, the particular teacher to whom a child happens to be assigned, the particular peer group he happens to become intimate with, the characteristics of his parents and siblings, accidents producing nervous system injuries, illnesses leaving neural defects, all of these impinge on the developing organism and interact with his current status. Such influences, e. g., characteristics of parents and sibs, are only partially correlated at best with the characteristics of the child. This means that motivation to learn fluctuates somewhat unpredictably and exposure to various kinds of learning is somewhat unpredictable. Both lead to unpredictability of future learning and thus to an uncertain future repertoire.

Biological development also does not proceed at the same rate for all structures nor for all individuals. Those who arrive at sexual maturity early tend to be taller than their age mates at that time, but achieve shorter adult height. There is a possible genetic basis for differential growth rates that

would account for reduced correlations between present status and future development. Thus it is not possible to rule out unevenness in biological development as at least a partial cause of the findings of Anderson and Roff. There is a seeming randomness in both the biological and psycho-social substrates that leads to imperfect predictions of future status.

2. Desirable human characteristics tend to be positively correlated with each other. This principle is particularly evident in unselected samples from the entire population. For example, in an American or Western European population the correlation between height and intelligence is approximately .25. There is evidence (Husen, 1959) that this relationship is not genetically determined but that it may be determined prenatally. As another example, the ability to make simple perceptual discriminations is positively correlated with general verbal knowledge. Some of these positive correlations may be determined genetically, some by the psycho-social environment, and some by biological "accidents." Whatever the explanation may be, however, the principle is important for a theory of intelligence.

Summary. This chapter introduced a behavioral definition of intelligence that goes beyond the simple statement that intelligence is what intelligence tests measure. The behavioral repertoire that is called intelligence and that is sampled under controlled conditions by intelligence tests, develops out of biological, including genetic, and psycho-social substrates, but without independent assessment of these substrates it is not possible to make inferences about them from a test score.

From this definition it follows that there are no qualitative differences among intelligence, aptitude, and achievement, but there are quantitative differences along three separate dimensions. These are the breadth of the repertoire,

its age, and its tie or lack thereof to a specific educational experience. From these defined properties of the concept of intelligence and from some very general principles of learning, genetics, and development, testable hypotheses can be derived. These are presented in the next chapter.

References

Anastasi, Anne. Differential Psychology. The MacMillan Company, 1958, New York, 664 pp.

Anderson, J. E. The prediction of terminal intelligence from infant and preschool tests. 39th Yearbook, National Society for the Study of Education, 1940, Part I, 385-403.

Fleishman, E. A., and Fruchter, B. Factor structure and predictability of successive stages of learning Morse Code. Journal of Applied Psychology, 1960, 44, 97-101.

Fleishman, E. A., and Hempel, W. E., Jr. Changes in factor structure of a complex psychomotor test as a function of practice. Psychometrika, 1954, 18, 239-252. (a).

Fleishman, E. A., and Hempel, W. E., Jr. The relation between abilities and improvement with practice in a visual discrimination reaction task. Journal of Experimental Psychology, 1955, 49, 301-312.

Harlow, H. F. The formation of learning sets. Psychological Review, 1949, 56, 51-65.

Humphreys, Lloyd G. Nature and organization of human abilities. Yearbook of the National Council on Measurement in Education, 1962, Ames, Iowa.

Husen, T. Psychological Twin Research. Almquist and Wiksell, Stockholm, 1959, 153 pp.

Jensen, A. R. Patterns of mental ability and socio-economic status. Proceedings of the National Academy of Sciences, 1968, 60, 1330-1337.

Roff, M. A statistical study of intelligence test performance. Journal of Psychology, 1941, 11, 371-386.

Tuddenham, R. D. Soldier intelligence in World Wars I and II. American Psychologist, 1948, 3, 54-56.

Tupes, E., and Shaycoft, M. Normative distributions of AQE aptitude indexes for high school age boys. Technical Documentary Report PRL-TDR-64-17, Lackland Air Force Base, Texas, 11 pp.

Wesman, A. G. Intelligent testing. American Psychologist, 1968, 23, 267-274.

HYPOTHESES DEVELOPED FROM THE THEORY OF INTELLIGENCE

Lloyd G. Humphreys

April 1, 1970

HYPOTHESES DEVELOPED FROM THE THEORY

Hypotheses derived from the theory are presented in this chapter. The deductions are not tight because the theoretical statements are not quantitative. Quantitative statements can come only from more and better data, and the rigor of the deductions is not important if there is a consensus that the conclusions do indeed follow from the theory. At any rate, the check of theorem against data is the conclusive step in the enterprise.

It is obvious that many of the hypotheses are circular; i. e., the theory was derived from the data concerning intelligence tests and the nature of intelligence tests, and the "tests" of the theory had known outcomes at the time the hypotheses were derived. Certain ones represent predictions from the theory for which data are not available, however, and consequently represent better checks on its adequacy.

Three important classes of hypotheses will be discussed. One class includes effects on mean performance of groups. A second class includes effects on stability of individual differences. The third class includes predictions or concurrent inferences made from intelligence tests. Both of the latter classes involve effects on correlations, but in the second class the emphasis is on the stability of intellectual performances while in the third class the emphasis is on generality.

Mean Performance of Groups. A few of the hypotheses that follow from the theory are almost trivial, but are worth stating as an antidote to common psychological and lay thinking concerning the fixed nature of intelligence. It must also be remembered that changes in means will be represented by arbitrary scales of measurement with a mean and standard deviation based upon the performance of some reference group. Change is expressed in age or grade units, or in standard scores within age or grade groups.

When data are used to support hypotheses in this area, it must also be recognized that experimental controls are frequently lacking. Statistical control involving some variant of the partial correlational technique such as covariance analysis is never a complete substitute for the control obtained through random assignment of subjects to experimental groups. For one thing, measurement error reduces the accuracy of statistical control. Failure to measure an important component of variance is a second source of inadequate control. It is also possible to control too much variance statistically and, as it were, throw the baby out with the bath. Partialling reading comprehension measures out of relationships involving intelligence tests would be considered suspect by most investigators. There would be more debate concerning the partialling out of a measure of socioeconomic status from those same relationships. The presence of debate and the lack of objective answers on such issues indicates all too clearly the hazards involved. The lack of experimental control does not mean that research work should cease on important problems. It does mean that a careful investigator will be modest with respect to the conclusions he draws from his data.

1. Change will occur. The evidence here was referred to earlier. Isolated and deprived groups show progressive declines in intelligence. The population of the United States, as evaluated by military tests, has shown a progressive increase in intelligence. Scottish children between 1933 and 1947 showed an increase in intelligence, as measured by the group test administered on both occasions (Scottish Council for Research in Education, 1949). After equating the 1916 and 1937 Stanford-Binet individual tests by an inadequate methodology, however, the original investigators concluded that "real" intelligence, i. e., that measured by an individual test, had not risen. The present writer using published data and a more adequate methodology (Humphreys, 1970) has shown that the

individual test results were almost completely parallel to the group test results. The Scottish gain is not as large as the American gain, but the Scottish retest occurred immediately after the close of World War II. The children tested had not, by any means, had a normal Scottish educational experience.

2. Measurable change will occur only with the expenditure of substantial effort. The literature concerning the effects of various educational methods is pertinent here. Training experiments lasting up to one semester and involving an hour or less per day have little differential effect on performance. The effects of brief cramming or review sessions prior to taking an intelligence (or college entrance) test are consistently very small. Nationwide testing program sponsors advise students that cramming will do little good. Yet when a young man attends a preparatory academy full-time for a year, the increase in scores on tests of the College Board averages approximately 100 points on the three-digit scale (Marron, 1965). Data for the separate tests are presented in Table 1. Marron also found that some preparatory schools produced greater gains than others, but no attempt was made to explain these differences.

Census figures show that the educational level of our population has risen in each decade. These figures reflect a very substantial additional educational effort over the years between the two World Wars and may well be a primary causal factor in the measured increase in intelligence over that same period. Furthermore, there has been some decrease in the growth rate of years of formal education since World War II, and there has been a corresponding decrease in the growth rate of intelligence.

3. For a given level of effort there will be greater effects on young children than on older children. Growth curves of intelligence as a function of age certainly do show decreasing returns with increase in age, but this finding is not

quite to the point with respect to the application of special efforts to facilitate growth in intelligence with different age groups. There is also a problem with regard to the units of measurement used since there is a general consensus that either mental or educational age units decrease in size with increase in chronological age. If change were measured in these units, empirical findings would almost certainly be the reverse of those expected on the basis of the hypothesis. Change must be measured, therefore, in relative units such as standard score or classical intelligence quotient units. If the problems involved in the measurement of both effort and intellectual growth are solved, however, it should be easier to obtain change when the repertoire is small than when it is large.

4. Changes in intelligence are a function of the kind of intervening educational experiences. Exposure to the traditional academic curriculum, with attention to the problem of the learner's motivation, should be effective in producing change in intelligence. Techniques of instruction conducive to the formation of learning sets and generalization tendencies should also be effective in producing change.

A good many years ago, Lorge (1945) published data on the relationship between retest gains on an intelligence test and intervening educational experience. There are also good recent data published by a Swedish investigator for changes between 13 and 18 (Harnquist, 1968). Since different tests were used at the two age periods, Harnquist obtained canonical composites. The major comparisons involve the first canonical composite which has reliabilities of .943 and .952 for the initial and final measures, respectively. The metric differs on the two occasions, however, with the initial standard deviation being 10.10 and the final 8.37. There may be a small ceiling effect on the final canonical composite.

Table 2 summarizes two of Harnquist's estimates of gain in the major educa-

tional groupings suggested by him. Also included are data for regressions of scores based upon the estimated within groups reliability of the canonical composite computed by the present writer. Hernquist concluded that the gains computed from scores regressed in accordance with the estimated "true" stabilities of the tests were probably most valid. Gains based upon difference scores corrected for differences in metric represent the most conservative estimate of gain. Gains computed from estimated reliabilities are intermediate.

The present writer has little confidence that he has the last word on the most appropriate method of estimating gain from these data. It does not seem reasonable to use stability coefficients, either obtained or corrected for errors of measurement, since the experimental conditions affecting the means also presumably affect the stabilities. On the other hand, something other than correcting for a change in metric is in order. The intermediate values based upon reliability are more conservative than those based upon stability of measures over time. By any method, however, gains are differentially associated with the amount and type of intervening education. While gains for the higher groups may be somewhat attenuated by the ceiling effect, it is also evident that the differences among groups are not spectacularly large, and there is much overlap. Many other factors beyond formal schooling are obviously involved. Since subjects were not assigned to groups at random, caution is also indicated concerning attributions of cause. Any laboratory analogue, on the other hand, must be considerably less realistic than the present "experiment" which lasted for 5 years.

5. For intellectual growth there must be a continuous supportive psycho-social substrate. There is no magic key or no critical time for intellectual stimulation. Temporarily successful Head Start type programs will not be successful in the long run if the children immediately revert to the prior psycho-social

environment. There must be continuing exposure and continuing effort. The exposure can be readily manipulated by the society, but the effort required is that of the learner. Social effort that does not affect individual effort will not pay off. Current evidence concerning these issues is almost entirely lacking, but the obtaining of such evidence is one of the most critical research issues of our time. It is also a difficult research area.

6. Change is slower for intelligence than for more narrowly defined abilities. Intervention in a narrow area will produce more rapid and larger amounts of change than intervention in a broad area. Differential gains on the so-called aptitude and achievement tests of the College Board resulting from preparatory school experience are relevant. Table 1 presented earlier showed that there is greater gain in English and in Mathematics achievement than in Verbal or Mathematical aptitude.

7. There will be little decline in intellectual performance in the absence of clearly discernible biological deterioration. Since there is little forgetting of overlearned and continuously practiced skills, the repertoire should not shrink. Older data seemingly contradict this statement. The more recent and better controlled research, however, indicates that the well documented decline is the result of failure to control intergenerational differences in intelligence. Older research was entirely cross-sectional. Cohorts of different ages were measured at the same point in time. The more recent data (Schaie, 1965) involved measuring different cohorts at the same point in time, but an additional test administration was required of the same cohorts five years later. The analysis of variance allows one to estimate contributions to variance of age cohort and of aging with the result that the former is found to make the main contribution. A reanalysis of the same data by Wachwitz (1970) shows this phenomenon even more clearly.

8. The educational practices of a society will have an effect on the age at which intelligence levels off. For example, pushing up the age limit for compulsory education should increase intelligence. It is of interest, in this connection, that the 1960 revision of the Stanford-Binet accepts the reality of mental growth for people in general to a higher age level than earlier editions of the test. The change in occupational patterns from a concentration of persons in manual labor to an increased proportion in more intellectual occupations should have a positive effect also.

9. There are mean differences in intelligence among groups defined demographically. This proposition is one of the best supported in the psychological literature. With the exception of sex differences on certain tests which were constructed to minimize such differences, all sorts of demographic variables show differences on intelligence tests without resort to Ns of astronomical size. Race, section of country, rural-urban, location, education of parents, education of examinee, school attended, level of teachers' salaries, etc., etc. all show differences. Interpretation of these differences is another matter, however. With adequate experimental controls an analysis of variance design could lead to estimates of percentage contributions to variance of psycho-social and biological substrates for the particular fixed levels of the independent variables studied. Results from a fixed variable design would hardly qualify as earth shaking in their implications for the heredity-environment issue, but in the absence of experimental controls conclusions with respect to percentage contributions are better characterized as meaningless rather than as limited in generality.

10. Among adult representatives of groups demographically defined it will be difficult to overcome existing differences. This proposition is independent of the attribution of degree of importance to psycho-social and biological substrates,

or within the latter to genetic versus acquired biological differences. Psychosocial deficits are not quickly and easily compensated for. Change takes place slowly.

11. There will be genetic differences among members of demographically defined groups whenever the definition of group accompanies some degree of segregation of gene pools. These differences will vary in size and sign of the difference from one of the very large number of biological characteristics to another. To take a concrete example, Negroes will be superior to Caucasians on some characteristics, inferior on others. The summation of the effects on developing intelligence from the entire gamut of biological characteristics will also show a race difference, simply because it is inconceivable that the algebraic summation of the effects of a very large number of partially segregated independent causal factors would be zero. On the basis of present data it is not possible to specify either the size or sign of this overall difference though it is certainly smaller than present observed differences in performance.

12. The selective breeding experiments with lower animals, such as those by Tryon (1929, 1940) could, with adequate controls, be replicated for high and low intelligence groups in the human. While this experiment will probably never be done, and with good reason, it is still useful to suggest the hypothesis. A summary of the controls necessary to reproduce the results with lower animals serves to make explicit the fallacies in the thinking of those persons who place great weight on social class or caste in human society.

The experiment starts with upper and lower groups of subjects selected from the tails of the distribution of intelligence. High subjects are mated only with high and low with low; all average subjects from the first generation are discarded. Subjects in the next generation are again measured and offspring who do not

meet the standards of their parents are ruthlessly discarded. After about a dozen generations of highly selective mating and discarding of unwanted offspring, rats show two distributions of maze running ability with very little overlap. The genetic substrate for intelligence in the human is probably more complex than the genetic substrate for maze running in the rat, for one thing there are more chromosomes in the human, so that many more than a dozen generations would be required to separate bright and dull groups an equal amount in the human.

Since the necessary conditions for the experiment are so greatly at variance with human breeding patterns, even in relatively highly stratified societies, there is no justification from this hypothesis for an assumption of large, fixed differences in genetic substrates among existing social classes and for the use of this reasoning as a basis for a highly stratified class society. For example, there is a common saying among conservatives that any revolution that abolished existing social classes would soon result in their reestablishment. While this seems to be true historically, and while it is reasonable psychologically as well, it overlooks two important factors: the new superior class would be composed of different people than the old, with many coming from the lowest social class; and the offspring of the new class would be inferior to their parents, just as the present class that currently is in a power position in a highly stratified society is inferior to their parents who were in turn inferior to theirs. That is, without both selectivity in mating and the ruthless discard of inadequate offspring, an initial genetic difference between persons of high and low achievement will diminish progressively in their descendants.

Stability of Individual Differences. Hypotheses in this area are mainly concerned with changes in the rank order of individuals over time. Time is not, of course, the effective variable, but in the absence of control of type of experience

or growth time is the appropriate dependent variable. Research that will pin down the factors that occur in time that produce the instability should have very high priority.

It will also be noted that many hypotheses are parallel to those in the mean performance of groups section. It seems reasonable that changes in rank order of individual differences will accompany changes in the mean status of groups.

1. Stability coefficients will always be smaller than reliability coefficients. Change is inevitable. While this generalization will be modified in subsequent theorems by such variables as age of the subjects, amount of time involved, and intervening experience, raw change is the primary phenomenon. It is a phenomenon, furthermore, with which psychologists concerned with prediction have not dealt in any systematic, comprehensive way.

2. Stability over time is a function of the age of the subject. With increasing age there is greater stability. This follows from the increasing size of the intellectual repertoire with increasing age and the relative size of increments to that repertoire as a function of age. John Anderson phrased the principle in terms of the characteristics of the part-whole correlation, assuming that increments were uncorrelated with the base at the beginning of the period. While his data were congruent with the latter assumption, it is not necessary to make that assumption in "deriving" the hypothesis. A correlation between increment and base that is lower than unity after correction for attenuation is a sufficient condition. Some degree of unpredictability of future learning or development is required, but not complete unpredictability.

It is well known that correlations between infant and early grade school tests of intelligence are approximately zero. This has traditionally been explained as due to a difference in functions measured by tests at the two time periods. This

explanation is unnecessary since change will take place rapidly starting with the very small infant repertoire. The data are almost precisely what would be predicted if the tests were measuring the same function. The only discrepancy between prediction and actual outcome, if it is real in the sampling sense, is between an expected small positive correlation and those obtained (Bayley, 1949) small negative correlation.

The degree of instability of intelligence and the increasing degree of stability with age, are well shown in the intercorrelations of mental ages obtained in the Harvard Growth Study. Data for boys are shown in Table 3 and data for girls in Table 4. One can also see in these tables some evidence for a period of increasing instability around the period of sexual maturity. This secondary instability appears earlier in the data for girls than for boys.

3. Instability over time is as characteristic of physical traits as of intelligence. While there is a psycho-social substrate for height and weight, it is reasonable to believe that the genetic substrate for height and possibly weight is relatively more important than for intelligence. Change in these characteristics is shown in Tables 5 and 6. Height is clearly more stable than weight and both are more stable than intelligence, but all show the same pattern of intercorrelations.

4. The amount of instability is a function of the amount of time between test and retest, holding age constant. The continuous addition of uncorrelated or lowly correlated increments to an initial base results in more and more change in the rank order of individuals with the passage of time. Data previously presented in Tables 3 to 6 confirm this hypothesis.

5. Change is more rapid with narrow than with broad functions. Other things being equal change should be more rapid in verbal or quantitative ability alone

than in intelligence. A prime example of this hypothesis is the learning of a motor skill. The intercorrelations of trials, or blocks of trials, all obtained during a single learning session, show the same pattern of instability found for intelligence and height over a period of several years. Changes in the rank order of individuals obtained in the course of half a day for a very narrow, rapidly acquired disposition are comparable to those obtained over a period of several years for a much broader, more slowly acquired disposition.

Stability coefficients for the College Board tests for the period from September to March were presented in Table 1 along with the gains made by students in a preparatory school. The aptitude tests show greater stability than the achievement tests. The former sample broader and older repertoires than the latter.

6. Change is a function of the intervening, psycho-social substrate. With respect to intelligence there should be more change in individual differences for students in an academic curriculum than in a skilled trade curriculum. There should be more change among a group of professional men than among a group of skilled workers. Change should also be dependent upon avocational interests. In general, the greater the opportunity to add to the intellectual repertoire, the greater should be the shift of individual differences as a function of the amount of time the exposure continues.

Harnquist (1968) presented regression coefficients for the several groups of subjects studied, but with standard deviations made available (1969) these can be converted to correlations. The within group correlations for the four major categories of type of education are presented in Table 7. Within group standard deviations are also shown. The results are in the expected direction, but they are also equivocal. The two groups whose experience has presumptively been less academic have larger standard deviations which might alone produce the higher

correlations obtained. There is, however, no applicable correction for restriction of range of talent. An interpretation involving restriction of range of talent, on the other hand, depends upon equal units of measurement in the several parts of the scale which the possibility of a ceiling effect makes suspect. Again, as with the mean gains, it can be said that the differences are not dramatic and that better control of intervening experience than that afforded by type of schooling will be necessary to test the hypothesis more precisely.

7. Degree of incentive to learn or strength of motivation present in a group will be positively associated with amount of change. Students in a highly competitive academically oriented educational institution will show more change in rank order of individual differences than will students in a more placid environment. It is possible that persons in a free, fluid society will show more change than persons in a highly structured society in which position is dependent on class or caste.

There appears to be no available evidence concerning this proposition. On an anecdotal basis, there may be more early stars that crash, and students that bloom late, at colleges such as Reed and Oberlin than in state colleges. An investigator would, of course, have to control range of talent for any work in this area.

There are data on amount of change in rank order of grade averages in a large state university over the four year time span (Humphreys, 1968), but there are presently no comparative data from other types of institutions. The sheer amount of change is sufficiently impressive, however, to give inferential support to the theorem. Intercorrelations of independently computed semester averages are shown in Table 8. It can be inferred that the changes in grades parallel, to some extent at least, changes in measured academic abilities.

8. The stability of individual differences in intelligence from person to person among a set of related persons is nonzero, but less than the reliabilities of the measures. This proposition, furthermore, follows from the influences of both the psycho-social and biological substrates. All substrates are involved in determining individual differences in intelligence and, except for monozygotic twins for whom there are no genetic substrate differences, all substrates differ among sets of related persons. The much discussed regression from parent to child, or from child to parent, for example, depends upon a finding of less than perfect correlations between parents and children and nothing more. Attribution of cause to the genetic substrate without independent assessment is barred here just as it is in interpreting the I. Q. of an individual. Parents and children have different childhood environments, the children themselves have different functional environments within the family, and different genotypes may interact with similar environments in a very dissimilar fashion.

The genetic interpretation of family resemblances does have one advantage over an environmental interpretation in that the degree of resemblance expected can be set with at least a modest degree of precision. The degree of precision must be called modest, however, because for psychological characteristics there is some degree of assortative mating, and the heritability coefficient is less than unity. Information is lacking as to the number of generations of assortative mating there has been and whether the same degree of resemblance between parents held in times past as in the present. Estimates of heritability of intelligence also vary from about .80 at the top to substantially lower values. By fixing either the genetic correlation, arising from assortative mating, or the degree of heritability, hypotheses involving a range of correlation coefficients can be tested.

9. Regression from initial standing to final standing in intelligence is

toward the mean of the identifiable subgroup of which the individual is a member. For the entire range of talent, without differential intervention in terms of intervening experience, the stability coefficient for intelligence will be less than the reliability coefficient. As a result subgroups defined by score on the initial test will regress more toward the population mean than would be expected on the basis of measurement error alone. Different forms of intervention may accelerate, retard, or reverse the expected regression toward the population mean and will involve instead the subgroup mean.

An example of the importance of this theorem is available in the folklore of higher education. It has been said that the graduates of superior colleges are no more superior then they were as entering freshmen. This allegation — firm data are lacking — is typically used to belittle the quality education claims of such colleges and places the emphasis on initial selection of the student body. On the basis of the present hypothesis, however, an institution that prevents the expected regression is doing a superior educational job.

It would not be difficult to obtain data concerning this issue. The College Board aptitude tests and the Graduate Record Examination aptitude tests are sufficiently similar that one could be quite confident concerning equipercentile conversions based on a random sample of applicants for college admissions. Comparison of pre and post test results for a variety of types of institutions would then be possible. There is one difficult matter that interferes with a complete assessment: the expected regression in the population in the absence of differential intervention is unknown.

The present hypothesis is intimately related to current social problems such as integrated education and admission of marginally qualified students to college, but the proposition is not sufficiently precise at the moment to make the needed

predictions. Certain extreme cases seem clear. A marginal student who quickly fails will not profit. A marginal student who is only slightly marginal and who survives should profit. Presumably each student should be pushed hard intellectually, but it is also possible to push too hard. But what is the result if the student is kept in a generally superior learning environment by means of special sections or differential standards of evaluation?

A partial answer to some of these questions is furnished by a reanalysis of the data in the Coleman report (1966). Using partial correlation techniques to control for variables such as socio-economic status, McPartland (1969) has shown that integrated classrooms seemingly increase the academic performance of Negroes while integrated schools having segregated classrooms do not. Similar studies need to be done on the academic performance of Caucasian children in integrated schools and integrated classrooms.

The efficacy of the several components of a superior learning environment is also unknown. In addition to faculty and facilities such as libraries and laboratories it is probable that the peer group itself is very important. If peers are important, the important ones would be the functional peers, or the significant peers, not merely those who happen to attend the same institution. In large universities particularly there are large numbers of functional peer groups having very diverse characteristics. Measurement of the characteristics of the functional peers, which Astin (1965) has done on an institutional scale, should provide very useful information.

Validities of Intelligence Tests. Test validities are usually described by correlation coefficients just as are the stabilities of individual differences. When the time interval between test and criterion is the critical variable, parallel hypotheses result. In these cases hypotheses in this section are presented with a

minimum of discussion. More attention will be given hypotheses concerned with the generalizability over content of inferences drawn from scores on intelligence tests.

1. The extent to which predictive validities of intelligence tests will decrease with the passage of time is a function of the age of the subjects. With increasing age there is less shrinkage of the validity coefficients.

2. The extent to which validities of intelligence tests will decrease over time is a function of the amount of time that intervenes between test administration and the accumulation of criterion information. Prediction of college grades, semester by semester, would seem to be an appropriate setting to test this hypothesis. The data obtained, which show that the problem is more complex experimentally than it appears superficially to be, are presented in Table 9. The predictive validities (Humphreys, 1968) fall off very nicely in accordance with the hypothesis. The postdictive validities (Humphreys, 1970) show that there has been a change in the rank order of students' academic abilities as a function of the educational experience, but the correlations for junior and senior grades are not as high as they should be if only changes in abilities were involved. While the hypothesis is supported, the amount of change was overestimated from the predictive validities alone.

3. Validity coefficients change more for narrow than for broad functions. Wechsler-Bellevue intelligence quotients should show a less steep gradient of validities than a college admissions test, since the Wechsler test represents a broader gamut of abilities than does the typical college admissions test. Empirical support for this hypothesis can be obtained from the postdictive study discussed under 2 above. Table 10 contains a comparison of correlations between the "aptitude" and advanced test sections of the Graduate Record Examination and semester grades. There is clearly more change in the correlations for the

narrower tests (Humphreys, ibid.).

4. The gradient of validities over time is a function of the content of the intervening psycho-social substrate. There should be a steeper gradient for intelligence tests in a highly academic curriculum than in a skilled trade curriculum.

5. The gradient of validities over time is a function of the degree of motivation present in the group. Size of validities in a highly competitive academic institution will shrink more than those in a more placid environment.

6. Gradients of predictive validities of intelligence tests are accompanied by similar gradients of postdictive validities. The gradients are not necessarily identical in shape, but age, time, and intervening experience will have similar effects both forward and backward in time.

7. Intelligence tests have a broad spectrum of concurrent and predictive validity coefficients. The intelligence test is broad, covering verbal, numerical, figural, and pictorial items requiring a wide range of types of responses such as association, comprehension, induction, deduction, memorization, etc. on the part of the examinee. Furthermore, desirable qualities are positively correlated. As a result it is difficult to find a criterion measure in the full range of talent for which an intelligence test does not have a positive nonzero validity.

8. It follows from 7 above that differential validity of narrow aptitude tests is difficult to establish in the full range of talent. The restriction of range associated with passage through the educational hierarchy affects the general factor primarily so that differential validity patterns are more readily observed in restricted populations such as college undergraduates. Validation studies in the military enlisted population support strongly this proposition.

9. Even though the validity spectrum is broad the very highest validities are obtained in educational settings. Test content is more like the academic

curriculum content than that of other common learning experiences. Within the educational setting the highest validities are obtained with criteria that have the most overlap in content with the test. Prediction of later reading comprehension and proficiency in arithmetic are higher than predictions of spelling accuracy. Music, art, and athletics are even less intellectual, in the present sense of that term, than spelling. Correlations with grades in foreign language courses stressing the spoken language are also low and, by the same token, the performance in foreign language training is not very intellectual.

10. There are many important criteria that are not predicted highly by scores on intelligence tests. An analysis based upon transfer principles is a reliable guide to the expected size of these correlations when test and criterion reliabilities are held constant. For example, leadership, sales, and manipulative criteria are not predicted well by intelligence tests.

Psychologists have been able to rationalize low correlations with the latter two criteria, but the first has been difficult to accept. Acceptance is made difficult by common beliefs concerning the nature and importance of intelligence and the importance of leadership behavior in our society. Correlations are frequently computed in a very restricted range of talent when leadership is involved, but this is only a partial explanation of their small size. When the same sample of officers is sent back to school for either officer or technical training, correlations with school grades become substantially higher than those previously obtained with rated officer effectiveness. In such comparisons, of course, the range of talent in intelligence is constant.

11. Theory is not now and will not in the foreseeable future be an adequate basis for the use of an intelligence test in a new situation or with a new population of examinees. Accurate use of a test requires a regression equation or an

equivalent actuarial table. It is not sufficient to decide that a test will be correlated with a particular criterion. Making predictions concerning individuals or groups requires precise information concerning errors of estimate, slopes of regression lines, and intercepts of regression lines. In spite of some 60 years of use of intelligence tests, furthermore, the amount of information required to use tests properly is still quite inadequate. The common definition of intelligence as a fixed general capacity along with the ease of making inferences from this interpretation is partially responsible for this state of affairs.

A case in point is the controversy concerning the use of intelligence tests for the "underprivileged." Typically, this boils down to a question concerning the use of tests for American Negroes. A consistent finding, though one not as broadly documented as it should be, is that for periods up to about one year the same regression equation can be used for members of both Negro and Caucasian groups for the prediction of a variety of criteria. Within this body of data there are some small exceptions to this generalization, chiefly with regard to the intercept of the regression of the test on the criterion, but the sum of these small intercept differences does not favor the Negro.

The naive environmentalist who accepts the common definition of intelligence as some entity inside the person may be dismayed by the above empirical findings, but they are quite reasonable from the point of view of the present theory. The intelligence test predicts later intellectual performance whether that performance be another test or a socially desirable criterion. It does this just because both occasions sample overlapping intellectual repertoires. The amount of overlap and the rapidity of change are functions of the variables previously discussed in this chapter.

12. A necessary empirical basis for concluding that low on-the-job validi-

ties, as opposed to high training validities, demonstrate that the job and training situations are functionally different involves both a predictive and a concurrent validity for measures of the same disposition. A low long range predictive validity and a high concurrent validity show that the people in the sample have changed. A low concurrent validity for an intelligence test, when the intelligence test was highly correlated with early training criteria, along with a significantly higher correlation for a test of some other disposition, is a necessary condition for concluding that training and job ability requirements are indeed different.

There have been many claims that on-the-job criteria have little relationship to intelligence. As a matter of fact some writers have gone so far as to claim that this is a nearly universal phenomenon. An implicit assumption basic to the claims that have been made to date is that abilities are fixed. Once this assumption is questioned, the controls that no one previously considered become essential.

13. Early training success is not a criterion of the degree of importance that it has assumed in test validation. The first 6 hypotheses concerned with validity are sufficient grounds for this assertion. In the absence of ability to predict changes, for many selection purposes retention or turnover has many attractive characteristics for criterion purposes. In deciding between early training success and retention as criteria questions that must be faced, among others, are the following: how much change takes place, how rapid is the change, how large are training costs, how much capacity for training is available or can be obtained, what are the characteristics of fast learners that slow learners would replace, what are the differences if any between asymptotic performances of slow and fast learners?

The preceding discussion does not presuppose that man is infinitely trainable

or that individual men are indefinitely trainable, but in the absence of information concerning capacity, which is not furnished by any aptitude test, one can not stake everything on initial training success. The only solution lies in more and better research.

14. The lowering of standards of initial selection for a group will result in lower final performance even though the time span between selection and performance is sufficiently long to reduce validity coefficients to near zero. This hypothesis is based on a previous one to the effect that change within a subgroup given special treatment is about the mean of that subgroup rather than about the population mean. Since the present hypothesis is a secondary one based in turn upon an earlier hypothesis it must be stressed that it is highly speculative.

Although speculative, this hypothesis is needed as an antidote to a different and probably overoptimistic inference from drastically reduced long term validity coefficients: namely, that initial selection does not matter. For example, in the well publicized World War II unselected group of pilot trainees (DuBois, 1947), if training standards had been reduced in line with the input, would the mean performance in the air of the group after training have been appreciably lower than the performance of control groups even though correlations with on-the-job criteria were essentially zero? There are no data concerning this question, but it should have high priority in an applied research program.

References

Astin, A. W. Who Goes Where to College? Science Research Associates, Inc., 1965.

Bayley, Nancy. Consistency and variability in the growth of intelligence from birth to eighteen years. Journal of Genetic Psychology, 1949, 75, 165-196.

Coleman, James S. et al. Equality of Educational Opportunity. U. S. Department of Health, Education, and Welfare, Office of Education, Washington, D. C. Government Printing Office, 1966.

DuBois, P. H. The Classification Program, Report No. 2. Army Air Forces Aviation Psychology Program Research Reports, 1947, Washington, D. C.

Harnquist, K. Relative changes in intelligence from 13 to 18. Scandinavian Journal of Psychology, Vol. 9, 1968, 50-82.

Harnquist, K. Personal communication. Unpublished manuscript.

Humphreys, Lloyd G. The fleeting nature of the prediction of college academic success. Journal of Educational Psychology, 1968, 59, No. 5, 375-380.

Humphreys, Lloyd G. Footnote to the Scottish survey of intelligence. British Journal of Educational Psychology, (to appear).

Humphreys, Lloyd G. A post diction study of the relationship between ability tests and undergraduate semester grades. Unpublished manuscript.

Lorge, I. Schooling makes a difference. Teachers College Record, 1945, 46, 483-492.

Marron, Joseph E. Special test preparation, its effect on college board scores and the relationship of effected scores to subsequent college performance. Office of the Director of Admissions and Registrar, U. S. Military Academy, West Point, N. Y., November 1, 1965.

McPartland, J. The relative influence of school and of classroom desegregation on the academic achievement of ninth grade Negro students. Journal of Social Issues, 1969, 25, 93-102.

Schaie, K. W. A general model for the study of developmental problems. Psychological Bulletin, 1965, 64, 92-107.

Scottish Council for Research in Education. The Trend of Scottish Intelligence. 1949, The University of London Press, London, 151 pp.

Tryon, R. C. Genetic differences in maze-learning ability in rats. 39th Yearbook, National Society for the Study of Education, 1940, Part 1, 111-119.

Tryon, R. C. The genetics of learning ability in rats: preliminary report. University of California Publications in Psychology, 1929, 4, 71-89.

Wackwitz, John H. Abilities as a function of age: an alternative to levels of performance. Psychological Bulletin (to appear).

Table 1

Gains on College Board Scores as a Function of Preparatory School Attendance

Score	N	September		March		Gain	Stability Coefficient
		$\bar{X}$	S	$\bar{X}$	S		
Verbal Aptitude	714	471	89	528	85	57	.81
Mathematics Aptitude	715	532	99	611	94	79	.83
English	649	458	82	540	93	82	.69
Intermediate Mathematics	610	497	89	629	100	132	.76
Advanced Mathematics	251	494	85	620	96	126	.74

Table 2

Gain as a Function of Type of Intervening Schooling

Type of Schooling	N	Initial Mean	Final Mean	Corrected Gains		
				Retest Regressed	Reliability Regressed	Standardized Diff.
Compulsory Level	1518	36.38	39.08	4.99	6.23	6.70
Vocational	946	39.46	40.23	6.72	7.50	7.80
Lower Secondary	558	44.49	43.40	8.58	8.13	7.96
Gymnasium	1194	49.90	47.00	10.39	8.55	7.86

Table 3

Intercorrelation of Mental Ages of Boys

at Various Chronological Ages (First Test)

	8	9	10	11	12	13	14	15	16	17	18
8		721	712	747	729	657	598	648	652	556	444
9	721		751	721	714	696	634	615	609	583	509
10	712	751		816	769	704	726	738	699	604	543
11	747	721	816		859	787	745	810	802	736	638
12	729	714	769	859		854	778	786	806	775	732
13	657	696	704	787	854		864	785	770	780	754
14	598	634	726	745	778	864		839	778	750	765
15	648	615	738	810	786	785	839		868	778	744
16	652	609	699	802	806	770	778	868		848	788
17	556	588	604	736	775	780	750	778	848		828
18	444	509	543	638	732	754	765	744	788	828	

Table 4

Intercorrelations of Mental Ages of Girls

at Various Chronological Ages (First Test)

	8	9	10	11	12	13	14	15	16	17	18
8		730	719	761	735	661	661	719	696	603	549
9	730		746	744	774	757	705	698	723	704	607
10	719	746		812	820	794	788	784	756	709	710
11	761	744	812		884	832	804	841	837	787	730
12	735	774	820	884		881	841	846	857	844	8.7
13	661	757	794	832	881		871	823	830	837	821
14	661	705	788	804	841	871		865	812	817	830
15	719	698	784	841	846	823	865		903	839	837
16	696	723	756	837	857	830	812	903		912	857
17	603	704	709	787	844	837	817	839	912		900
18	549	607	710	730	819	821	830	837	857	900	

Table 5

Intercorrelations of Standing Height of 275

Girls at Various Chronological Ages

	7	8	9	10	11	12	13	14	15	16
7		987	980	957	920	897	887	866	836	810
8	987		989	969	934	914	904	882	850	824
9	980	989		986	954	927	909	881	844	814
10	957	969	986		979	947	911	865	816	780
11	920	934	954	979		974	923	855	790	747
12	897	914	927	947	974		964	887	810	763
13	887	904	909	911	923	964		961	901	860
14	866	882	881	865	855	887	961		974	948
15	836	850	844	816	790	810	901	974		989
16	810	824	814	780	747	763	860	948	989	

Table 6

Intercorrelations of Weight of 275

Girls at Various Chronological Ages

	7	8	9	10	11	12	13	14	15	16
7		890	880	835	810	793	755	773	744	732
8	890		920	896	871	856	825	812	771	759
9	880	920		932	906	882	840	818	773	756
10	835	896	932		958	936	892	842	777	755
11	810	871	906	958		967	921	866	790	762
12	793	856	882	936	967		954	892	816	775
13	755	825	840	892	921	954		944	880	839
14	773	812	818	842	866	892	944		953	916
15	744	771	773	777	790	816	880	953		965
16	732	759	756	755	762	775	839	916	965	

Table 7

Stability as a Function of Type of Intervening Schooling

Type of Schooling	N	Initial S. D. (Within Groups)	Final S. D. (Within Groups)	Correlation (Within Groups)
Compulsory Level	1518	8.44	7.00	.67
Vocational	946	8.29	6.92	.67
Lower Secondary	958	7.48	5.40	.56
Gymnasium	1194	7.35	5.08	.56

Table 8

Intercorrelations of Independently Computed Semester Grade Averages for a Constant Range of Talent

(N is approximately 1600 for each correlation,

	I	II	III	IV	V	VI	VII	VIII
I		556	456	439	399	415	387	342
II			490	445	418	383	364	339
III				562	496	456	445	354
IV					512	469	442	416
V						551	500	453
VI							544	482
VII								541
VIII								

Table 9

Comparison of Predictive and Postdictive Validities of College Aptitude Tests

	I	II	III	IV	V	VI	VII	VIII
Predictive								
ACT English	345	278	226	236	236	222	216	160
ACT Math	279	189	171	171	145	162	156	121
Postdictive								
GRE Verbal	349	308	255	268	251	218	213	163
GRE Quant.	348	333	311	291	275	205	170	146
Corrected to Freshman Range of Talent								
Predictive								
ACT English	40	35	27	25	22	22	24	20
ACT Math	40	30	25	23	20	20	18	15
Postdictive								
GRE Verbal	42	40	31	33	31	28	28	21
GRE Quant.	43	43	38	37	34	26	22	19

Table 10

Comparison of GRE Aptitude and Advanced Test Validities

(Correlations are computed within groups defined by Advanced Test and then aggregated; Aptitude Test validities differ somewhat from those in Table 9 which were computed within College and sex.)

Restricted Sample

	I	II	III	IV	V	VI	VII	VIII
Verbal	297	285	262	281	275	256	223	195
Quantitative	270	246	209	253	217	215	203	6
Advanced	286	304	347	359	336	343	316	258

Corrected to Freshman Range of Talent

	I	II	III	IV	V	VI	VII	VIII
Verbal	37	36	33	36	35	33	28	25
Quantitative	34	31	27	32	28	28	26	20
Advanced	36	38	43	45	42	43	40	33

THE PSYCHOLOGICAL TEST

The status of measurement in a discipline is intimately related to the status of both research and theory in that discipline. Little sophisticated research on electrical phenomena could be done until measuring devices such as voltmeters, ammeters, and ohmmeters were developed. Research was necessary in order to develop the measurement devices, with the first "measurements" being simply presence or absence of the phenomenon, but the devices also led to better research and theory. Furthermore, as measuring devices became more sensitive, the range of experiments possible was extended. The ability to measure in microvolts may represent as important a step as the one from presence or absence of voltage to the first voltmeter.

Importance of the Test for Theory. It is not generally recognized, however, that the type of measurement available in a discipline also affects research and theory. The psychological test, for example, represents a type of measurement device found infrequently if at all elsewhere in the sciences. It is essential to understand the nature of tests if one is to understand experimental or observational correlates of tests or the theory that is developed from those correlates. A scholastic philosopher can define intelligence in the absence of measures of intelligence, but a psychologist qua psychologist can not do so.

The preceding statement does not assume the necessity for an operational definition of each term in a scientific theory. A direct measure is not required for each theoretical construct, but there must always at some point be a return to data. The data in a discipline, in turn, depend on the measures. A theory that requires measures which do not exist and which can not be developed is not testable and theories which are not testable are not acceptable scientific theories.

It is the thesis of this book that most theories of intelligence are not scientific theories in the above sense. This particular section will develop

psychological test theory in some detail to lay the groundwork for the thesis and as a basis for developing a theory of intelligence that is congruent with the experimental and observational correlates of measures of intelligence.

Example of an Ordinal Scale. The first characteristic to be discussed is that the psychological test furnishes only an ordinal scale of measurement. Suppose that an investigator wishes to measure the number of English words known by a particular population of people, e. g., high school students. He could define a population of English words by means of the unabridged dictionary and devise a method of sampling words at random from that population. (Note that a well defined population of test questions is not ordinarily available to the test constructer which creates a problem to be illustrated by later examples.) The investigator obtains a list of 100 words by his sampling method and presents this list to a random sample of the population of people in whom he is interested and asks for definition.

The answers given must be scored and to score in anything like an objective, replicable fashion a scoring key must be developed. Will the key demand word for word definitions more or less as they appear in the dictionary? Will the test author accept instead approximate definitions, including reasonably close synonyms? Or will he be satisfied if the word is used in a phrase or clause in a fashion that conveys generally the meaning of the word, indicating at a minimum that the subject has seen the word used somewhere?

Clearly the number of words that are counted as correct will depend on the key which in turn will determine the estimates of the total size of the vocabularies of the subjects. The latter computation is made simply by multiplying the number correct on the test by the ratio of number of words in the dictionary to the 100 sampled by the test, but the figure obtained is relatively meaningless. Depending nature of the test key, an individual's estimated vocabulary can vary

tremendously.

Is anything gained by converting the test from the original open-ended, or recall, version to an objective test format such as multiple choice? The objective test will be easier to score and with care the scoring can be accomplished with zero error, but subjectivity in the writing and interpreting of the key for the recall version has been pushed back into the selection of the misleads. By selecting misleads that capitalize upon fine nuances of meaning, the test can be made very difficult, and the subjects may appear to have restricted vocabularies. On the other hand, by selecting misleads that require only the grossest of discriminations, the test can be made quite easy.

It would be easily possible by any method of test construction to obtain three vocabulary tests with quite different distribution characteristics in the same population of subjects. For example, one could obtain means and standard deviations approximating those in the following table for each of three randomly selected 100-item tests with just a little trial and error.

	Test A	Test B	Test C
Mean	25	50	75
Standard Deviation	20	25	20

It is also reasonable to assume that there will be no gaps in the distributions of scores and that most of the possible range of scores will be represented on each of the tests. Furthermore, in large samples of subjects, say 1000 or so the distributions, whatever their shape will appear quite regular. One does not need third and fourth moments of the three distributions, furthermore, to draw inferences about the shapes of the three distributions. Test A and C are skewed, though in opposite directions, while Test B, though probably symmetrical in distribution, is more platy-kurtic than the normal distribution. These inferences all follow from the relationships of the standard deviations to the means.

Characteristics of Ordinal Scales. The linear intercorrelations of the three tests may be quite high--even a moderately well constructed test of 100 items for a high school population should be quite reliable--but these correlations can not be as high as the test reliabilities would allow. Some high scoring people on Test A, each with different total scores, will have identical scores on Test B and even more of them will have identical scores on Test C. A similar finding for low scoring people on Test C will be noted in comparing their scores on Test B and A. Such cases arise from the differential skews of the three distributions and regressions are necessarily curvilinear. In general the number of units by which any two scores differ in one distribution will be different than the number of units by which scores comparable in rank order differ in either of the other distributions.

It is also clear that ratios of scores computed for any one of the three tests are uninterpretable. The number 50 is twice 25, but getting 50 items right vs. getting 25 items right does not have the same meaning for each of the tests. Zero, furthermore, is quite a common score on Test A, is much less frequent on Test B, and is very rare if it occurs at all on Test C. A score of zero on any test would not indicate that the subject had a vocabulary of zero length. The accidents of sampling from the population of words are involved, but more importantly the arbitrary decisions of the test constructer serve to make a score of zero meaningless with respect to absolute size of vocabulary. The only information about a score of zero furnished by the test is that it is smaller than one.

The preceding characteristics of the three test scores define ordinal scales of measurement. The information furnished is basically rank-order information even though the numbers used have the appearance of an interval scale. The rank orders of subjects inferred from the numbers are not identical for the three tests, even allowing for measurement error, because the differences in skew will produce tied ranks on one test that are not tied on another. If we utilize the rank-order

information, however, and convert the obtained, raw scores to standard scores by means of a monotonic nonlinear transformation, i. e., by working through the percentile ranks, we can increase the linear correlations among the three tests as compared to linear raw score correlations.

Other examples will be presented to develop the argument in more detail and in more generality, but for the present the vocabulary test example can be taken on faith to represent the general case. Distributions of test scores are arbitrary. Tests furnish rank-order information only. Since a normal distribution has a number of desirable statistical properties, it is recommended that raw scores on tests be converted to normal distributions by means of the nonlinear transformation involving percentile ranks in a random sample of some defined population. When this has been done, the scale of measurement is said to have been normalized, but it is still ordinal. It has not been converted to an equal interval scale simply by means of the transformation. The choice of the normal curve conversion is a matter of convenience not of scientific necessity or conformity to natural law. If convenience dictates a different type of distribution, e. g., quartiles, deciles, or centiles for the converted scale, a different type should be used.

The fact that measurement with a test is ordinal is only the most obvious characteristic of this form of measurement. It is far from being the most important. It will become clear later that ordinal measurement has little effect on reliability or validity. Most of the inferences from test scores that are barred by the lack of interval or ratio scales are relatively unimportant and substitutes are generally available.

Possible Functions of Multiple Items. In the example of the preceding section tests of 100-items were assumed. One hundred different words were selected at random from an unabridged dictionary and subjects were asked to define these or to select the correct alternative from a list of misleads. This suggests a property

of the test that is of the utmost importance.

Tests are typically compsed of multiple items or "hurdles" with the subject behaving in some fashion with reference to each item. In measuring ability the performance required is either right or wrong, but in measuring personality or interest answers are frequently yes or no, like or dislike, etc. The total score on the test is also typically a linear combination of the scores on the items and, in many tests, weights for each item are either zero or one. It is not an essen tial feature of the test that the scoring be dichotomous, although dichotomous scoring is found very frequently. It is also not essential that the combination of items be linear, but nonlinear combinations can be dismissed from this discussion on grounds that such combinations are used only infrequently for research purposes and rarely if at all in standardized tests. The theory to be developed will assume a linear combination of dichotomous items for the sake of convenience and for wide-spread applicability, but the theory is directly applicable to all other types of item scoring with only minor modifications. Major modifications would be necessary, however, to adapt it to nonlinear combinations.

For the person steeped in traditional measurement theory the first hypothesis concerning multiple items is that they are required for purposes of reliability. All measurement is subject to some degree of measurement error. A scientist or engineer frequently makes multiple independent readings of his measures and uses the mean of these as his best estimate of the "true" value. Does the test differ in any respect from the need to take multiple measurements to reduce error?

There is, indeed, a difference in the practices of the engineer and of the psychologist. The person using a psychological test, when he wishes to increase his precision of measurement, repeats the whole test or uses parallel forms of the original test to obtain his sample of measures for which the mean is the best estimate of the subject's "true" score. The total test score is considered the

measure, not the score on an item. It is true that increasing the number of items in the test generally has the effect of increasing the reliability of the total score, but increasing reliability is not the primary function served by use of multiple items.

A second hypothesis concerning the function of multiple items is that they are required to furnish a scale more nearly approximating a continuous scale of measurement than does the dichotomous item. There are occasions, however, when it is not merely desirable but necessary to add together multiple items each of which is measured on a continuous, equal interval scale, i. e., total score composed of multiple items may be required even though the items are not dichotomous. Again, multiple dichotomous items do furnish a scale approximating a continuous measure, but this is secondary to their primary function.

The Important Function Served by Multiple Items. The principal function served by multiple items is best seen as a contrast between test theory per se and traditional measurement theory. In the ordinary measurement of height it is reasonable to assume that each measurement operation for each person measured includes a true score component and a random error component. This is the starting point for classical measurement theory. The variance of obtained scores includes the variance of true scores and the variance of error. Correlations involving the obtained scores are a function of the covariances with true scores and the variances of obtained scores. From this basis such statistical concepts as the standard error of measurement and correction of correlations for attenuation by measurement error are readily developed.

Classical measurement theory is not, however, readily applicable to the test. Efforts to use classical theory over the years, furthermore, may have hurt test development as much as it has helped. The major departure from classical theory arises from the necessity to start with a definition of item score that differs

significantly from classical theory. An item score is composed, as Loevinger has discussed (1954) most fully, of three distinct parts: score on the trait or disposition (d) in which the test constructer is interested, systematic nonerror noise or bias elements or factors (b), and random error (e). The important effect of using multiple test items is to minimize the effects of the numerous nonrandom factors that are subsumed under the label of noise.

To return to the vocabulary example, a high school student may have encountered a word in his recent reading for which he obtained a definition and when this word was encountered on the test he answered it correctly. The word may be difficult in general and the student's general vocabulary level low, but he obtained an extra point in his total score for a nonrandom reason independent of his general level of vocabulary competence. There are many such examples. Some words are encountered more frequently in science than in the humanities, in pulp magazines than in school books, or in certain neighborhoods or social levels than in others, and so on. By taking a large sample of words such effects, although still present, can be balanced off against each other, and the more general disposition to know the meanings of words will be measured with greater validity.

In this connection it is instructive to look at item intercorrelations for some standard ability tests. In tests that are quite homogeneous both with respect to difficulty level of the items and the subject matter of the test, item intercorrelations with a mean as high as .20 are not common in the full range of talent and occur quite rarely in special groups who are restricted in range of talent. While it is not easy to assess the random error component of variance in an item--memory for item content makes suspect a repeated measures design and parallel items are not easy to construct--it is probable that nonerror or bias factors are a major contributor to item variance for most of the items that appear psychological tests.

The vocabulary example suggests an alternative designation of nonrandom noise or bias in test items as item sampling error. This is indeed one source of bias, but the equivalence breaks down in two important ways. In many, many cases there is no defined population of items from which to sample, though the use of item selection error could get around this difficulty. A more important difficulty, however, is that certain bias factors are intrinsic to psychological items. Every test item has a particular item format, a time limit or work limit, a set of directions. In addition, each examinee has a different background of knowledge, skills, sets, and other experiences. All contribute variance to a test item. Reasoning is necessarily measured with verbal symbols, numerical symbols, or figural materials. Words that appear in a vocabulary test occur with differential frequency in different kinds of reading material. The use of noise or bias suggests unwanted or even uncontrollable, which is desirable, and the use of systematic indicates that the behavior measured is lawful, which is also desirable.

Just as a weed is any plant growing where it is not wanted, systematic noise or bias includes any factor or element appearing where it is not wanted. One man's noise, for one purpose, becomes another man's primary mental ability, for another purpose. But unlike weeds, a great deal of systematic bias is impossible to eradicate.

The Correlation Between Test and Criterion. A little algebra may be helpful at this point. Let there be n items in the test and let d, b, and e represent the disposition the test constructer desires to measure, the bias factors, and random error, respectively. The correlation of the test of the disposition with a criterion measure of the disposition is given by the following:

$$r_{xy} = r_{(x_i + \ldots + x_n)y} \text{ and each } x_i = d_i + b_i + e_i$$

$$r_{xy} = \frac{\Sigma\, Cd_{iy} + \Sigma Cb_{iy} + \Sigma Ce_{iy}}{S_y \sqrt{\Sigma S^2_{d_i} + \Sigma S^2_{b_i} + \Sigma S^2_{e_i} + \Sigma\Sigma\, C_{d_i d_j} + \Sigma\Sigma\, C_{b_i b_j} + \Sigma\Sigma C_{e_i e_j} + \Sigma\Sigma C_{d_i b_i} + \Sigma\Sigma C_{d_i e_i} + \Sigma\Sigma C_{b_i e_i}}}$$

If error is truly random, then all covariance terms involving e will drop out. Psychometricians have typically been willing to make this assumption. Moreover, if the b terms were specific to each item and independent of the disposition score on the item, they would have the functional characteristics of error and covariance terms involving b would also drop out. While it is not difficult conceptually to assume orthogonality of disposition and bias factors, the assumption of specificity of bias to each item is almost always false.

It is also reasonable to assume that the noise factors are unrelated to the criterion measure. (This can be considered true by definition.) With these considerations in mind, formula 1 can be rewritten as follows:

$$r_{xy} = \frac{\Sigma C_{d_i y}}{S_y \sqrt{\Sigma S^2_{d_i} + \Sigma S^2_{b_i} + \Sigma S^2_{e_i} + \Sigma\Sigma\, C_{d_i d_j} + \Sigma\Sigma\, C_{b_i b_j}}}$$

In the best of cases bias factors are minor sources of variance of total score on the test (denominator) and make zero contributions to covariance with the outside variable (numerator) currently of interest. In the worst of cases the nonrandom bias variance of the test is entirely noise from the point of view of the aims of the test constructer, and the only nonzero terms with criteria involve sources of variance other than the one the test is supposed to measure.

By basing the total score on many items, it is possible to build up the validity of the test for a particular disposition even though any one item has only a small component of that disposition. The secret is to spread item selection over as many bias factors as possible so that any one bias factor runs through a minimum number of items. The goal, though frequently unattainable, is to make the bias factors specific to items. Even when it is impossible to keep the bias covariance terms near zero in the denominator, the scattering of this

variance among many bias factors will avoid the situation in which the total score is a better measure of some other disposition than the one intended. Many items, therefore, are a necessary though not a sufficient condition for building up the variance of a particular disposition in the total score on the test. A basic misconception concerning the original choice of items will result in the test constructer measuring something, with greater and greater precision as he continues to add items, that he does not wish to measure.

INTERRELATIONSHIPS OF HOMOGENEITY, RELIABILITY, AND VALIDITY

The concept of homogeneity of a test does not appear in classical measurement theory. Homogeneity with respect to content is an issue only in those situations in which multiple items are used. The statement made earlier that multiple items did not serve the same primary purpose as multiple measures in physics or engineering, but did have a secondary effect on reliability is important in this connection. Many psychologists are confused on this issue. Homogeneity interacts with both reliability and validity, but must not be confused with either.

Homogeneity and Reliability. Kuder-Richardson homogeneity coefficients are frequently called reliability coefficients. Under certain restricted circumstances, it is true, one can obtain a reliability estimate from a measure of the homogeneity of the test, but it is essential that the investigator keep the distinction, and the conditions, clear in his own mind and in his writing.

The Kuder-Richardson formula best used to estimate reliability is the following:

$$r = \frac{n}{n-1}\,\frac{s_x^2 - \Sigma p_i q_i}{s_x^2}$$

Only the number of items, the difficulty levels of the items, and the variance of the total score on the test are used. (This is algebraicly equivalent to the approach to homogeneity of a set of measures by means of the analysis of variance which Hoyt (1940) suggested.) The variance of the total score is, in turn, a function of the number of items, the item variances, and the item covariances. These parameters do not have a one-to-one relationship to reliability defined as the correlation between repeated measures or between repeated parallel measures.

The difference, and the relationship, between homogeneity and reliability can best be shown by writing out the formula for the repeated measures correlation as a function of the relationships involving the items. (The prime refers to the repeated or parallel item or test total score.)

$$r_{xx'} = r_{(x_1 + x_2 + \ldots + x_n)\,(x'_1 + x'_2 + \ldots + x'_n)}$$

$$r_{xx'} = \frac{\Sigma C_{x_i x'_i} + \Sigma\Sigma\, C_{x_i x'_i}}{\sqrt{\Sigma S^2_{x_i} + \Sigma\Sigma C_{x_i x_j}}\ \sqrt{\Sigma S^2_{x'_i} + \Sigma\Sigma\, C_{x'_i x'_j}}}$$

If item intercorrelations are zero, the right hand term in the numerator disappears as do the right hand terms under each of the radicals. The reliability coefficient is then completely a function of the item reliabilities and can vary from zero to one. Furthermore, one can conceive of a test in which these conditions would be rather closely approached. A scored biographical data blank, for example, could contain items that were essentially uncorrelated with each other, but the reliability of answering an individual item would be very high.

As the intercorrelations of items within and between tests approaches the correlations between the paired items, the Kuder-Richardson homogeneity coefficient approaches the reliability coefficient of the test. For the two to be equal in conception the item difficulties would all have to be the same. Otherwise covariances are necessarily somewhat attenuated. In practice, this latter condition can be ignored since the formula does take out the variance due to the main effect of difficulty level, and variations of difficulty level within the normal range have only a slight, biasing effect on the interaction between persons and items which is the essential term determining the homogeneity of the test.

The interaction between reliability and homogeneity is more clearly seen if Formula 5 is rewritten to make explicit the assumption that the retest or parallel measure is identical with the first; i. e., test variances are equal and intercorrelations within tests are equal to intercorrelations between tests.

$$r_{xx'} = \frac{\Sigma C_{x_i x'_i} + \Sigma\Sigma\, C_{x_i x_j}}{\Sigma S^2_{x_i} + \Sigma\Sigma C_{x_i x_j}}$$

The right hand quantities in the numerator and denominator are identical. The left hand quantities represent the ratio between paired item covariances and variances. With many items in a test the right hand quantities will generally be much larger than the left hand ones; the homogeneity of the items, in other words, typically makes a larger contribution to reliability than the reliability of the paired items. The test constructer by narrowing the focus of his test, i. e., by redefining the disposition in which he is interested to make it coincide with an important source of nonerror noise, can step up test reliability very easily. To suggest that this may be undesirable may seem strange to those imbued with classical measurement theory. Why should not the ratio of true score variance to total variance be maximized? The answer is, of course, that an increase in reliability is not worth the price if the disposition which the psychologist seeks to measure is redefined in the process of test construction to make it less useful psychologically.

As a matter of fact, the positive steps in test construction that follow from the concept of disposition, nonrandom noise, and error contributions to item variance make it difficult to achieve high reliability with a limited number of items. The variance of noise or bias factors must be spread around as widely as possible. The more successful the test constructer is in his efforts, the lower will be the item covariances. He can compensate for this effect only by increasing the number of items in the test.

Homogeneity and Validity. No one administers a test, however, simply to obtain reliable information of some sort about a person. Tests are administered in order to make inferences about behavior: inferences about jobs, school, military assignments in applied work, or inferences about functional relationships involving a particular disposition in more basic research. Validity is a shorthand term used to cover the inferences that can be drawn from a score. Validity

coefficients stated in terms of item characteristics were presented in Formulas 1 and 2, but a simpler one will now be more convenient. This one is stated in terms of the items, without regard to their components, and their relationship with any outside variable, y.

$$r_{xy} = \frac{\Sigma C_{x_i y}}{S_y \sqrt{\Sigma S^2_{x_i} + \Sigma\Sigma C_{x_i x_j}}}$$

Formula 7 shows that, item validities being equal, there is a premium placed on low homogeneity. Item covariances occur only in the denominator. High reliability which comes about through an interaction with homogeneity is indeed a misleading goal. Only in case certain subsets of items in a heterogeneous test have zero correlations with the criterion does it pay to increase the homogeneity of the test and obtain the concomitant increase in reliability. When all items are related to the outside variable, by keeping item intercorrelations low, the variance of the test score will be kept low and reliability will be kept low, but the size of the validity coefficient increased. Such reasoning is completely compatible with expectations based upon multiple regression theory, but it does require qualification of the classical theory concerning the relationship of reliability to validity.

Reliability and Validity. Classical theory states the relationship of reliability to validity in the correction for attenuation.

$$\hat{r}_{xy} = \frac{r_{xy}}{\sqrt{r_{xx'}}\sqrt{r_{yy'}}}$$

This formula is applicable to test theory only in cases where reliability is changed by the addition or, with appropriate changes in the formula, subtraction of items of the same type as the originals. Whenever items are discarded

selectively with others retained, an increase in reliability may accompany a decrease in the validity of the test. Increasing the reliability of a test by doubling its length with exactly comparable items will increase the test's validity. Increasing the reliability of a test by item selection procedures will not have a predictable effect on the test's validity.

The same assumption, i. e., adding exactly comparable items, must be made in estimating the reliability of a test of a different length than the original, but the importance of the assumption in this case is better known. It may be instructive, however, to apply it to the hypothetical situation described earlier in which item intercorrelations in the original test are zero.

When item intercorrelations are zero, test score reliability is more or less the mean of the item reliabilities. When the length of the test is doubled by the addition of exactly comparable items, item covariances are no longer zero. The assumption of comparability means that each item in the original test now has itself or a parallel version of itself in the test of increased length. The new test no longer has zero homogeneity.

Minimum Requirement for Homogeneity. Even though completely uncorrelated items would be best to maximize the correlation between a test and an outside variable, such a set would not be considered to measure a psychological disposition. It is here that the concept of the homogeneity of the test is required. Various indices of a disposition of psychological interest just ought to have something in common. If the disposition of glass to shatter or beams to snap under the stress were measured in a fashion analogous to the psychological test, the various items would be correlated just as the items that measure height or weight in the other physical analogues to the test are correlated.

If the reality, or even the necessity, for some degree of homogeneity is accepted, it does not necessarily follow that homogeneity should be as high as

possible within the limitations of obtaining a measuring device furnishing a nearly continuous score. (If a test of height were given very reliably, maximum homogeneity would result in a U-shaped distribution. Some degree of spacing of item difficulties, with the resultant decrease in item covariances, is necessary in a test for height to discriminate among the examinees.) Some degree of homogeneity is expected, but the degree is optional. The degree depends both upon the psychological facts, i. e., the extent to which behavior is dispositional as opposed to situational, and upon the breadth of the disposition that the psychologist wishes to measure. He may be interested in measuring intelligence, or perhaps something even broader than intelligence, or at the other pole in measuring the fluency with which four letter words beginning with s can be evoked.

Let the rule be that any set of positively intercorrelated items can be added together. Such composites produce a psychologically meaningful total score, particularly if the items are intercorrelated at about the same level. It does not matter whether this level is low or high. If one is trying to measure some disposition that is very broad, and in consequence each item may contain only a very small portion of the variance of the disposition, it will be necessary to plan on using many items widely scattered in order to dissipate the many possible sources of nonrandom noise. That other factors will contribute to the total score is of no consequence. As a matter of fact, the larger the number of these the better since this will tend to keep the contribution of each small. The restrictions that item intercorrelations be at about the same level is necessary in order to avoid giving undue weight to a particular bias factor. It can, however, be relaxed if this is carefully done. The bias factors must themselves be evenly distributed.

The Goal of High Homogeneity. In contrast to the preceding rule, those who set high homogeneity of the test as their goal, imply that only those items can be

added together that have the very highest level of intercorrelations. If any given test can be broken down into subsets of items whose intercorrelations are a little higher than the cross correlations between subsets, the original test is "impure" or heterogeneous and new tests should be constructed as defined by the subclusters of items.

The reader who believes the rule that any set of positively intercorrelated items can be added together is ambiguous--after all there are many possible levels of item intercorrelations and thus many possible tests--should ponder the ambiguity in the high homogeneity rule. How high is the highest possible level of intercorrelations? When does the correlation between two different items become sufficiently high that it should be considered the correlation between parallel forms of the same item? If this rule is pushed to the extreme, does it not mean that the ultimate in homogeneity is reached when one reaches a small set of items that are essentially parallel forms of each other?

In discussing Binet's interest in multiple intellectual functions and the development of the Binet scales of intelligence, Guilford (1967) concluded that Binet's decision to use a single score for the totality of his items (mental age) was completely incongruous. In the light of the present discussion Guilford's conclusion is simply incorrect. One can accept multiple factors both of the Thurstone sort and of the Guilford sort, which appear to be narrower than the Thurstone primary mental abilities, along with a general factor without any logical or psychological difficulty. (See Humphreys, 1962, for a fuller discussion of this issue). Ability items are positively intercorrelated to varying degrees. High intercorrelations determine narrow factors; moderate intercorrelations determine somewhat broader factors; low intercorrelations determine a general factor. It is completely reasonable for an investigator to measure with a single test the factor or complex of factors that produce the lowest positive correlations

among broadly distributed items.

It is important to realize that the argument here is empirical, not logical. Good characteristics as defined socially of the human being tend to be positively correlated. The most disparate abilities, with abilities used in the most general sense, such as correlations between clerical and mechanical abilities, or between information about farming and about social sciences, are positively correlated. Psychological dispositions of the abilities sort are also positively correlated with physical measures such as height and weight. Terman's gifted children were healthier, wore fewer glasses, etc., on the average, than other children.

This tendency for "all good things to go together" is much more marked when one samples from men or women in general in a given cultural group than when samples are drawn from more restricted ranges of talent. Even within samples of college students enrolled in the most highly selective institutions correlations still tend to be positive though occasionally negative values occur which can generally be explained in terms of sampling errors. Negative correlations in a restricted population do occur, but there is frequently a sampling explanation. Highly selective univeraities can not play big-time football without having separate standards for athletes and nonathletes. Correlations between athletic abilities and intellectual abilities will be negative in such mixed groups.

Arguments Pro and Con. One argument advanced against the broad test is that "purer" tests are better than more complex tests on grounds basically of scientific esthetics. Here there is a difference in point of view as to what constitutes a pure test. Guilford's factor pure tests are seen by the present writer as inextricably complex. Tests of high homogeneity that measure one of the "aptitudes" in Guilford's structure of intelligence reflect simultaneously variance introduced by all of his three dimensions. Such tests are like the physical analogue in which weight was measured by having each subject lie down in a uniform manner at the end

of the lever: scores were highly homogeneous, but reflected both height and weight in an unknown combination.

A second argument against broad tests, and an almost convincing one, is that all of the information in the most complex test is basically available in a large number of highly homogeneous tests of the Guilford sort. Potentially also, information is lost by moving from many tests to a smaller number of broader tests. There are two different counter arguments to this point both of which are matters of feasibility. It would be very difficult to motivate examinees to sit through and work well for the amount of time necessary to administer 120 tests each sufficiently reliable to justify a separate score. It is also very difficult statistically to obtain stable weights for 120 measures for the various sorts of inferences in which psychologists are interested. It is not an exaggeration to estimate that the number of cases required would run into the tens of thousands for each outside variable considered. This estimate has a statistical basis in the formula for the standard error of a beta weight and an empirical basis in the ubiquity of positive intercorrelations among items and tests.

It is even possible given optimum Ns for weighting purposes, little if any information would be lost by the use of broad tests carefully constructed (Humphreys, 1962). Tests of the analogue to the main effects in an analysis of variance for each of Guilford's dimensions might well furnish the same information as the 120 tests representing the Cartesian product of his dimensions. The 120 different combinations of those dimensions become a source for selection of items and a guide for distributing noise factors as widely as possible in this conception, but not a mandate to construct 120 tests for assessment purposes.

ILLUSTRATIONS OF TEST CHARACTERISTICS BY MEANS OF PHYSICAL ANALOGUES

It is possible to develop physical analogues to the test that help to clarify the principles that have been presented. These principles will also be developed more fully as the various physical analogues are discussed.

Behavioral Test of Height. A carpenter is asked to make a series of standards in the form of an inverted L with the only specifications being that the uprights will all differ from each other and that they will cover the range in height of adult men. It is not essential that the horizontal bar be at right angles with the upright, and the essential specifications are checked perceptually only. Each standard is given a separate designation, perhaps a number. A sample of men is drawn from a population; each man is confronted with each of the standards in turn in a uniform manner; and each man is given a score representing merely the number of times the tip of the horizontal bar touched his head when an attempt is made to pass it over his head with the upright being placed vertically on the ground.

If a very large number of standards is constructed initially, it should be possible to select from the larger group a smaller set having any specified distribution of item difficulties. (This statistic is readily computed: the number of heads hit by a standard divided by the total number of men in the sample measured provides a statistic varying from zero to 1.00 with high values representing "easy" items.) If 9 standards are selected having difficulty levels ranging from .10 to .90 by steps of .10, the distribution of total scores for height will be rectangular in shape; i. e., symmetrical but highly platykurtic.

It is not difficult to see how the shape of the distribution of total scores can be inferred from the distribution of item difficulties. The easiest item has a difficulty index of .9, i. e., 10% of the sample fails the easiest item. Their total score on the test will be zero. Another 10% fail the item having an index

of .9. While 20% fail this item, one-half of the failing group passed the easier item. Thus 10% will have a score of 1. In a similar fashion the 10% who passed the most difficult item passed all easier items. Since this is the 9th standard in order of difficulty, only 10% of the sample will have a score of 9.

With 10% of the sample at each score, the distribution is rectangular.

It is instructive to make a table in which the items are placed in a horizontal array in order of increasing difficulty and the subjects are placed in the vertical array in order of increasing size of total score. (In this example only, all subjects having the same score can be represented by a single tally.) The result, in Table 1, is a triangular matrix of tallies which defines what has become known as a perfect Guttman scale. No man fails an easier item after having passed a more difficult one. When the number of tallies in a column is counted and divided by the number of people, the difficulty level of the item is the result. When the number of tallies in a row is counted, the total score on the test is a result. (A percentage score on the test is sometimes obtained by dividing the total score by the number of items in the test, but the number of items and the zero point are quite arbitrary.)

It is of interest that the product-moment intercorrelations of the items in a perfect Guttman scale form what Guttman has called a simplex matrix (Guttman, 1955). The simplex matrix indicates the presence of a single underlying function or factor when the successive variables differ in difficulty level, complexity, growth, or level of learning (Humphreys, 1960). The intercorrelations of the present example are presented in Table 2 for purposes of illustration. It should also be noted in connection with this example that the presence of a single factor is inferred from the form of the correlational matrix and the known differences in difficulty level of the items. The simple, one factor explanation cannot be ob-

tained by the application of the usual factor analytic methods. If squared multiple correlations are alternated with unities in the diagonal,

the 9 items will define four principal component factors (Humphreys, 1960).

What would the test constructer do if he wished to obtain a normal distribution of test scores for his measure of height? He would go back to his population of standards and select those having an appropriate distribution of item difficulties. Difficulties ranging from .96 through .89, .77, .60, .40, .23, .11, to .04 would produce a distribution of total scores that would be approximately normal. The mean would be the same as the mean of the rectangular distribution, but the standard deviation would be smaller.

The test constructer by appropriate selection of item difficulties can produce a distribution having any shape he desires. Wide variations in both kurtosis and skewness are possible. A test for the selection of basketball players can be produced having a tail at the upper end of the score distribution. After all, the coach is not concerned about making discriminations among college freshmen who are in the lower quarter, or even half, of the distribution of height. U-shaped distributions are possible though hardly useful. For a general purpose test, however, the test constructer does not worry very much about the distribution of item difficulties needed to produce a raw score distribution having a particular shape. Instead he takes what the accidents of item selection produce and converts the raw scores, by means of a nonlinear, monotonic transformation into a distribution of converted scores. As indicated earlier, it is frequently convenient that the shape of the converted score distribution be normal.

A very important reason why the test constructer does not worry too much about the selection of item difficulties in most cases is that psychological items ot readily scale in the Guttman sense. One reason why this is true is the

unavoidable presence of measurement error in each item. While the presence of the Guttman scale made possible the characterization of the shape of the score distribution from knowledge of the distribution of item difficulties alone, it was also necessary to assume error free measurement in order to obtain that scale. The measurement situation had to be highly standardized. Instructions to subjects were given to control posture; instructions to the test administrator controlled the nature of the surface on which the subject stood and the placement of the standards relative to the subject.

The Introduction of Measurement Error. This principle can be illustrated by returning to the test of height and the population of standards originally postulated. The only change to be introduced is that uniform conditions of measurement will not be specified. Posture will no longer be controlled. Neither will the measurement surface nor the placement of the standard be controlled. All of these will be allowed to vary at random from subject to subject and from item to item within a subject.

When nine items are now selected having the same equally spaced distribution of item difficulties as before, the distribution of total scores will no longer be rectangular, but instead will be unimodal. When subjects and items are tabled as before, many persons will be found who have failed an easier item after having passed a more difficult one. Something like the item data in Table 3 will be the result though Table 3 is schematic only if it is taken to represent more than 10 subjects. The number of subjects who fail easier items after passing, will be functions of the amount of error that has been introduced by the failure to standardize the measurement situation.

Limitations of the Ordinal 4 Scale. The usual statistic expressing the reliability of measurement of a test is the correlation between repeated tests or between parallel forms of the test. If the conditions of careful, standardized

measurement did produce a perfect Guttman scale, the correlation between test and retest, or between two separate tests having identical item characteristics, would be unity. In a larger group of items occasional reversals would be found under even optimum conditions of measurement and the reliability coefficient would only approach unity. There is no essential reason, however, why the reliability of the test of height should not be every bit as high as the reliability of the usual measurement of height. It is all too easy, however, to be careless with any scale of measurement, and it is probably easier to introduce error into the test than into the use of a physical scale of measurement just because there are more occasions with multiple items for error to occur. Without standardization of the measurement situation, as in the second example, reliability coefficients will depart substantially from 1.00.

There is also no essential reason why the correlation between the test of height and the criterion measure of height should not approach unity. Lack of uniform conditions for the test, as well as carelessness in the measurement situation for the criterion, will attenuate the validity coefficient of the test, but there is nothing intrinsic to an ordinal scale that produces a reduction in validity. The only inferences barred are those involving equal units or equality of ratios and the absolute zero. For most purposes to establish converted scores in a meaningful population of subjects provides useful though not full substitutes for the standard deviation (requiring equal units) and the mean (requiring an absolute zero) of the ratio scale.

Probably the most important type of inference barred by an ordinal scale is the characterization of the form of the functional relationship between a psychological disposition measured by the test, as the dependent variable, and some independent variable. There is no point in worrying about power versus log functions, example, if there is doubt concerning the equality of the units of measurement.

By making certain assumptions about the nature of human judgment it is frequently possible to get outside the limitations of the test as here defined and obtain equal units. Problems of scaling have been discussed thoroughly by Torgerson (1958). For present purposes it is sufficient to add that interval and ratio scales formed by such assumptions must be thoroughly and independently checked. Thus, the supposedly equal interval attitude scales of Thurstone and Chave (1929) do not have intervals that are equal independent of the attitudes of the judges who do the scaling. The lack of equality as a function of attitude of the judge is more marked for the equal appearing interval method of scaling than it is for the paired comparisons method, but it is not completely absent in the latter (see Edwards, 1957, for an extended discussion of these data). It is also true that a Likert type scale (Likert, 1932), which is clearly ordinal in the sense here described, is probably just as valid as a Thurstone scale (Edwards, 1957).

Determinants of Test Score Distributions. With respect to the shape of the distribution of test scores, two generalizations are possible at this stage of the development: (1) the variability of item difficulties is inversely related to the variability of the distribution of scores on the test, or is directly related to a change in the form of the distribution toward leptokurtosis. (2) The amount of error present in the testing situation is inversely related to the variability of the distribution of scores, or directly to a change in the form of the distribution toward leptokurtosis.

The second generalization above appears to be at variance with classical measurement theory. It is easy to prove in the classical theory or in the Loevinger variant of that theory that the variance of true scores or of disposition scores is always less than the variance of obtained scores. It is not always noted, however, that the preceding conclusions demand an interval scale of measure-

ment. For the test this means, when the same set of items is administered once carefully and once carelessly, that two different ordinal scales are the result. The set of items administered carefully will have the larger standard deviation, but the ratio of the variance of true scores to error will also be larger in that set. In contrast, when height is measured carelessly on the physical scale, the variance of the obtained measures is larger, and the ratio of true score variance to error variance is smaller, than when height is measured carefully on the same scale.

Another way of illustrating these principles is to write the standard deviation of the distribution of total scores on the test in terms of the item characteristics. The effects of dispersion of item difficulties and the introduction of error can be observed in the item statistics.

$$s_x^2 = \Sigma\, p_i q_n + \Sigma\Sigma\, C_{ij}$$

The largest contribution to total variance of the item variance terms is obtained when $p = q = .50$ for all items. The largest contribution to the item covariance terms, on the assumption that all covariances will be positive (all items are assumed to be measuring the same function), is obtained when $p = q$ for all items. Wide variation in item difficulties reduces the contributions of item variances and covariances. For most tests the covariance terms are more important than the variance terms just because there are so many more of them.

The effect of increasing the amount of random error in the measurement situation comes about by way of attenuation of the size of the covariance terms. Error decreases the size of correlations among obtained scores relative to true scores. The greater the amount of error, the smaller the size of the item covariance terms which reduces the size of the standard deviation of test scores. The variance of a test score distribution attenuated in size by the presence of error of measurement will contain a larger proportion of error variance relative to true score variance than will the larger variance of an error-free test of the

same number and distribution of item difficulties. The absolute size of the standard deviation is smaller, however, for the error-ridden test.

It is obvious also that the size of the standard deviation of test scores is a direct function of the number of items. The addition of items of any type, error-free or error-ridden, will increase the size of the overall standard deviation. The addition of even a single item to a test changes the scale of measurement.

The Introduction of Nonrandom Bias. In order to consider effects of nonrandom noise or bias it will be useful to construct another physical analogue to the test. A test constructer interested in measuring weight has a lever, a fulcrum, and a pile of big rocks. In a pilot study he finds a place for the fulcrum that will allow the typical rock to just about balance the typical male adult. (The use of mean rock and mean adult has been avoided to indicate that the pilot research does not have to be precise.) Again the rocks are each given an identification and each man in the sample is placed on the lever opposite each of the rocks in turn. The score is the total number of rocks raised in the air by the passive man. The test constructer has a partially incorrect theory about what he is trying to measure however, and carefully instructs his subjects to lie down on the lever with their bare feet precisely at the end and with their heads extending toward the fulcrum as much as necessary. (He has been thoroughly indoctrinated with the necessity for care in measurement.)

Proceeding as before, ten items are selected which produce a rectangular distribution and a perfect Guttman scale. Scores represent an unknown mixture of height and weight, but there are no data from the measurement operations alone that lead to this conclusion. As long as great care is taken in measurement, no man will fail an easier item after passing a more difficult one. The first generalization from this new example, therefore, is that systematic measurement of nonrandor

noise does not necessarily reveal its presence.

If the test constructer had instructed his subjects to lie down as described on a specified one-half of the items and to stand with their heels at the tip of the lever on the other half, scores would still represent an unknown mixture of height and weight, but there is a possibility that the presence of a second factor in the items could be detected. In this new example of measurement of systematic bias, there are two types of items which would be expected to show their differential similarities in their correlational patterns which would in turn determine two factors. Unfortunately, item correlations are affected by range of difficulties as well as by content (see Table 2 in this regard). Differences in item marginals may cloud the statistical differentiation between the two factors, but there is hope in this example in being able to show the presence of the two factors in the data. Even with careful measurement, when two factors are present in the items, a Guttman scale will not be obtained; some subjects will fail easier items after having passed more difficult ones.

Even if the two types of items are separated perfectly on the basis of information internal to the measurement operation, there is no statistical clue from the item data as to which is the better measure of weight. The proper identification of the function each cluster is measuring might be made intuitively from inspection of the items in the separate clusters, but external relationships represent a more dependable means of identification. Thus the factor analytic method is vulnerable on two counts: (1) the difficulties in factoring dichotomously scored items and (2) adequate identification of factors without pursuit of differential, external relationships.

If posture is allowed to vary from standing to sitting to lying down and if this variation occurs at random from subject to subject and from item to item, was nonrandom noise becomes random noise. The effects of error have been

described earlier. There is, however, an in-between condition which is more serious. If posture varies from subject to subject but not from item to item, one subject's score may represent a relatively pure measure of weight while another's identical score may represent a mixture of height and weight. Again, as in the first example, there are no internal clues. A Guttman scale can be obtained under such conditions, for example, but information external to the measurement operation is necessary in order to identify those subjects whose measures are valid measures of weight. Without separation of subjects this type of nonrandom noise would depress validity coefficients. If subjects could be separated, however, validity would be very high in one sub-group, quite low in another.

Increasing Complexity of Bias Factors. Although a number of test construction principles have been demonstrated by means of physical analogues, up to this point in the development there has been no precise, complete analogue for the most important reason for the use of multiple items described earlier. Before introducing such an analogue it will be useful to return briefly to that argument.

Tests require the subject to behave on each of a number of items. An underlying trait or disposition to behave in certain ways is inferred from the test score. Yet there are myriads of possible causes for behavior. Any one bit of behavior may reflect the underlying disposition only in small degree. Knowledge of any one word does not indicate very much about a disposition to know many words. This phenomenon was described in terms of an analysis of the test score into disposition, nonerror noise, and random error components. Individual test items generally contain much more variance from nonerror noise and random error than they do from the hypothesized disposition.

A physical analogue that will illustrate this property can again be devised. A test constructer without a tape measure hopes to measure height. He also has no rods long enough to meet the criterion used in the first analogue example.

He can construct items that will measure the length of toes, fingers, lower arms, lower legs, and head; various measures of width and depth of arms, legs, trunk, and head are possible; various circumferences can also be measured. Such items could be dichotomous, but might even be measured by tape or calipers on the physical scale of measurement.

The first principle to note for this analogue is that, if dichotomous, the items would depart radically from a Guttman scale. Many, many failures on easy items after passing more difficult items would be evident. Long fingers do not typically accompany a broad chest. Also, for a given number of items, the standard deviation of the test scores would be lower than in previous examples for items of similar difficulty levels and with equal amounts of error. The distributions would be unimodal, even with minimum error in the measurement operations, and with little variability of item difficulties. Systematic noise or bias of this type, which is typical of psychological tests, reduces mean item intercorrelations substantially. The net effect on the test score distribution is similar to the effect of error. With sufficient item heterogeneity the test constructer does not have to distribute item difficulties in order to have a useful ordinal scale. Item difficulties clustered closely around .50 will produce a U-shaped distribution with highly homogeneous items carefully measured, but the same distribution of item difficulties with heterogeneous items carefully measured will produce a unimodal distribution of total scores.

Criteria for Item Selection. If there were an objective external criterion of height available, (a) it would be simple to obtain multiple regression weights for the set of items, (b) but under these circumstances it would be unnecessary to measure height with a set of items. The analogue to the test is to combine these items selectively in a linear fashion on the basis of internal data alone.

An equally weighted linear combination of all of the items listed would undoubtedly produce a score that would be rather highly correlated with height. This score would probably be more highly correlated with height than would any one of the components, but some sub-set of these items might produce an even higher correlation. Clustering or factoring might be possible, subject to the reservations expressed earlier about the effects of disparate item difficulties, but this approach is far from simple. Depending upon the density of sampling of bodily measures one might find both finger and toe length factors, a long bone factor, a body width factor or factors, circumference factors, etc. If only one factor is to be used, it is probable that the long bone one is most highly correlated with height, but the test constructer working without this knowledge would have difficulty justifying this selection. From item data alone his grounds could only be intuitive, following inspection of the items, and the items obviously refer to bone length and not to full stature. Furthermore, it is highly probable that the entire set of items carries more information about height than does the long bone subset, so the problem of the test constructer is to bring in all of the useful information and exclude the useless information from his test.

This situation can also be viewed in terms of the factor analytic methods with particular reference to the problem of factoring in several orders. With the use of very large numbers of items as contemplated in this test of height, there would probably be several first order factors defined by anatomical location and by the dimension, i. e., length, breadth, or depth, measured. These factors would contain relatively little information concerning full stature; instead they represent mainly systematic noise. Factors more closely related to full stature would be found in higher orders. With many width and breadth measures along with the measures of length, the factor in the highest order would probably represent body volume while stature would appear at the just lower order. This complexity obviously leads to a problem of factor identification and a requirement for exter-

main functional relationships. In general, however, desired dispositions are not first order factors.

The test constructer can proceed with simpler statistical methods, such as item-total score correlations (internal consistency item analysis), than factor analysis but there are several problems here. An important one concerns the original item pool. If it contains substantial numbers of width and depth measures, item selection by means of the total score correlations will lead to a measure of body volume rather than of stature. Secondly, there are no criteria for deciding at what point to exclude an item from a test. If the original item pool is approximately correct, but if homogeneity standards are set too high, useful items will be excluded; if set too low items not contributing to the measurement of height, as distinguished from the correlated measures of weight, will be included.

Test constructers frequently use the steepness of age or grade curves for items administered to children as a criterion of item selection, but there are many functions that increase with increasing age. Such age curves have been commonly used in the development of tests of intelligence, but their use has again been dependent on the original item pool. Items from the pool not showing the expected relationship with age are discarded objectively, but many, many items that would have shown the expected relationships with age do not appear in the pool. For intelligence tests the choice of items for the pool has been based upon theory, tradition, and availability--and not necessarily in that order of importance.

Without recourse to an external physical measure of height the test constructer can only make use of the network of functional relationships involving his test with measures of other functions as a check on his item selection. The implications for items of the functional relationships involving total scores are quite indirect so that the change of a measurement construct by varying item

selection is a slow, arduous, and ambiguous task.

Summary of Multiple Item Function. It is hoped that the function of multiple items, as well as the difficulties inherent in their use, has become clear. The constructer of a psychological test has no physical measure to use as a criterion. He measures bits of behavior (items), each of which reflects the underlying disposition in which he is interested only to small degree. By adding the right items together he can build up the variance of the underlying disposition in the total score, but he needs to reduce random error and to spread nonrandom noise as much as possible. In the present physical analogue the test constructer is not interested in measuring finger or toe length as such. Instead his interest lies in their ability to give him a modicum of information about stature. If he includes too many measures of finger and toe length in his stature score, there will be too much variance present from factors in which he is not interested. Such bias may be present in sufficient amount to mask the information about stature. By bringing in as many indicants of stature as possible, and varying their distribution over the body as much as possible, the nonrandom noise while still present is minimized or spread over so many functions or factors other than height that the total score reflects height primarily.

For a psychological function such as verbal comprehension used in an earlier illustration the availability of a population of words from which to sample randomly, and in sufficient number, is a very important way in which to define the central function or factor that one wishes to maximize in the total score. However, item populations from which to sample are relatively rare; when items are invented, the defining of a population is at best arbitrary and at worst impossible. To distinguish between the nonrandom variance of the disposition and of noise involves a long term research operation. The most hopeful procedure if one is restricted to all evidence is factoring the intercorrelations of carefully constructed items

in orders beyond the first. Much more often than not, first order factors represent systematic noise rather than the disposition in which the test constructor is interested, while it is only in the second or higher order that he finds the construct he is seeking to measure.

The above reasoning represents a complete break with a tradition of test construction in psychology in which high item homogeneity has been an important goal. The traditional reasoning has been that, with high homogeneity, one could infer that the test was measuring a unique, unitary function or factor. With sufficiently high homogeneity, the test becomes a Guttman scale though this is rarely attained. Scalability of a universe of items, nevertheless, became a goal toward which to strive.

This tradition leaves undefined the question as to how high the degree of homogeneity should be. This represents a formal objection to the homogeneity model, but the primary objection represented in this discussion involves the nature of test items. Items necessarily involve several kinds of components. For certain purposes one of these can be labelled the primary disposition while others are necessarily nonrandom noise and random error. Guttman scales are unobtainable except under highly restrictive, even artificial, situations.

Radical solutions should always be entertained, including the possibility that separate Guttman scales should be constructed for each source of nonerror noise, while trying to hold constant all other sources of bias. What has been noise, in other words, becomes many tests. The feasibility of this solution is in serious doubt, however, in terms of the sheer number of tests that would result. Any estimate must be labelled a guess, but it is a guess conditioned by test construction experience of psychometricians generally as well as by psychological assumptions concerning the number of possible sources, or causes, of responses to items. The items in ability tests that are generally considered to be quite

homogeneous typically have mean levels of intercorrelations less than .20. An attitude measurement approximations to Guttman scales can be obtained with a few items having very diverse item popularities (defined statistically in a fashion similar to item difficulty) in which essentially the same question is asked with minor variations in wording. Correlations between attitude items and logical reversals of those items are not high. All in all, a guess that tens of thousands of tests would be the result is not out of line. It seems better to retain the concept of nonerror noise and to allow test constructers freedom to broaden or narrow it at will, and in accordance with scientific convenience, rather than to impose the goal of high homogeneity on all tests.

TABLE 1

Score Matrix Which Produces a Perfect

Guttman Scale

		Items									
		1	2	3	4	5	6	7	8	9	Test Score
	a	+	+	+	+	+	+	+	+	+	9
	b	+	+	+	+	+	+	+	+	0	8
	c	+	+	+	+	+	+	+	0	0	7
	d	+	+	+	+	+	+	0	0	0	6
ersons	e	+	+	+	+	+	0	0	0	0	5
	f	+	+	+	+	0	0	0	0	0	4
	g	+	+	+	0	0	0	0	0	0	3
	h	+	+	0	0	0	0	0	0	0	2
	i	+	0	0	0	0	0	0	0	0	1
	j	0	0	0	0	0	0	0	0	0	0
Difficulty Level		.9	.8	.7	.6	.5	.4	.3	.2	.1	

TABLE 2

Intercorrelations of the Items in a Perfect Guttman Scale

	1	2	3	4	5	6	7	8	9
1		.67	.31	.41	.33	.27	.22	.17	.11
2			.77	.61	.50	.41	.33	.25	.17
3				.80	.66	.54	.43	.33	.22
4					.82	.67	.54	.41	.27
5						.82	.66	.50	.33
6							.80	.61	.41
7								.77	.51
8									.67
9									

TABLE 3

Score Matrix (Schematic) in Which Measurement Error Has Been Introduced on Potentially Scalable Items

		Items									
		1	2	3	4	5	6	7	8	9	Teat Score
	a	+	+	+	+	+	0	+	0	+	7
	b	+	+	+	+	0	0	+	+	0	6
	c	+	+	+	0	+	+	0	0	0	5
	d	+	+	+	0	0	+	0	+	0	5
Persons	e	+	+	0	+	+	0	+	0	0	5
	f	+	0	0	+	+	+	0	0	0	4
	g	+	0	+	+	+	0	0	0	0	4
	h	+	+	+	0	0	+	0	0	0	4
	i	+	+	0	+	0	0	0	0	0	3
	j	0	+	+	0	0	0	0	0	0	2
Difficulty Level		.9	.8	.7	.6	.5	.4	.3	.2	.1	

Abstract

A methodology has been described and illustrated for obtaining an evaluation of the importance of the factors in a particular order of factoring that does not require factoring beyond that order. For example, one can estimate the intercorrelations of the original measures with the perturbations of the first-order factors held constant or, the reverse, estimate the contribution to the intercorrelations of the original measures from the first-order factors alone. Similar operations are possible at higher orders.

EVALUATING THE IMPORTANCE OF FACTORS IN ANY GIVEN ORDER OF FACTORING

Lloyd G. Humphreys, Ledyard R. Tucker, and Peter Dachler

University of Illinois, Urbana

One of us (Humphreys, 1962) has recommended hierarchical factoring of measures of human abilities for reasons connected with a presumed gradient of importance of factors in the several orders. One indication of importance is predictive validity. Broad tests have generally higher predictive validities than narrow tests. It is also very difficult empirically to find stable differential weights for a variety of criteria for very narrow tests.

Valid objections can be raised to the evaluation of the importance of factors based upon correlations with outside criteria, but by factoring in several orders and using the Schmid-Leiman transformation (1957) to obtain a hierarchical orthogonal factor matrix an internal criterion can be obtained. The contributions to common factor variance of the several factors can be computed and compared. It occurred to us, however, that an internal criterion that did not involve higher order factoring would be useful. Such a criterion is readily available.

Mathematical Development

We shall let R stand for any matrix of intercorrelations. The subscripts 0, 1, 2, etc. will designate original intercorrelations, intercorrelations of first-order factors, intercorrelations of second-order factors, etc. Matrices of rotated factor loadings (Harman's pattern matrices on the primary axes) will be symbolized by P, also with appropriate subscripts. Estimated matrices are designated by the circumflex. Thus, we have the following well known relationships:

$$[\hat{R}_0 - \hat{U}_0^2] = P_1 R_1 P_1' \qquad (1a)$$

$$[\hat{R}_1 - \hat{U}_1^2] = P_2 R_2 P_2' \qquad (1b)$$

An evaluation of the importance of the factors in a given order is obtained by replacing the unities in the diagonal of R with the estimated communalities of the factors and multiplying as before. We symbolize the new matrix, which represents the estimated intercorrelations at a lower order with the perturbations of the factors at the next higher order removed, in a fashion analogous to partial correlations. Thus we have:

$$\hat{R}_{0.1} = P_1 (R_1 - \hat{U}_1^2) P_1' \qquad (2a)$$

$$\hat{R}_{1.2} = P_2 (R_2 - \hat{U}_2^2) P_2' \qquad (2b)$$

Other matrices of interest can be derived immediately from the above. The direct contribution of the first-order factors, in contrast to the control of their effects, is given by the following:

$$[\hat{R}_0 - \hat{U}_0^2] - \hat{R}_{0.1} = P_1 \hat{U}_1^2 P_1' \qquad (3a)$$

$$[\hat{R}_1 - \hat{U}_1^2] - \hat{R}_{1.2} = P_2 \hat{U}_2^2 P_2' \qquad (3b)$$

These direct contributions of the factors can be designated as $\hat{R}_{0.2,3..k}$ and $\hat{R}_{1.3..k}$ also in a fashion analogous to partial correlations to indicate that the effects of higher order factors have been removed. Thus the entries in the matrix $\hat{R}_{0.2,3..k}$ indicate the contributions of the first order factors only to the correlations among the original measures. It also follows that $\hat{R}_{0.1,2,3..k} = \hat{U}_0^2$, which represents another example of the symbolism.

It has been suggested that low $\hat{U}^2$ values for factors may be obtained when there is substantial capitalization upon chance (Horn, 1966), or when rotations have been contrived to force data into a particular structure (Humphreys, 1967). Low values of $\hat{U}^2$ can also be obtained legitimately and objectively from the nature of the data. Whatever the basis may be, if the matrices of contribution to correlations obtained in the fashion of Formula 3a contain values close to zero, the first order factors in question can be considered relatively unimportant. A "real"

factor may make only a minor contribution to covariation. A similar statement may be made for $\hat{R}_{1.3..k}$ and second order factors.

It is possible to estimate communalities for factors in the several ways that one estimates communalities for the original measures, but under certain conditions squared multiples have much to recommend them. Multiple correlations, depending as they do on the entire matrix of intercorrelations, are more stable than many other communality estimates. Although squared multiples are lower bound estimates only in the population of observations and approach "true" communalities only as the population of measures is approached, they tend not to be seriously in error when the number of measures, or factors as in the present case, is moderately large and when the number of observations is much larger.

When the use of squared multiples is appropriate, it is unnecessary to factor in a higher order in order to use formulas 2 and 3 in evaluating factor importance. Thus the error of estimate variance, symbolized as S^2, can be substituted for $\hat{U}^2$ in formulas 2 and 3.

$$\hat{R}_{0.1} = P_1 (R_1 - S^2) P_1' \qquad (4)$$

$$\hat{R}_{1.2,3..k} = P_1 S^2 P_1 \qquad (5)$$

We can now, in turn, let $V = PS$, which leads to the following relationship:

$$\hat{R}_{0.2,3..k} = V_1 V_1' \qquad (6)$$

The matrix V contains the projection of the measures on the normals to the hyperplanes (Harman's reference vector structure). Consequently, these projections can also be used to evaluate the contributions of the factors. Small correlations with the reference vectors of measures of high communality indicate that the important factors are in a higher order.

Illustrations of the Procedures

In order to illustrate these procedures we turned to published data. The

Adkins and Lyerly analysis of reasoning tests (1952) contains sufficient numbers of factors in the first order to allow use of squared multiples as communality estimates. The same is true of the first-order factors in Cattell's study of fluid and crystallized ability (1963).

It is not feasible to present large estimated intercorrelational matrices. Other indices must be found. Distributions of the diagonals of the several matrices constitute one compact way of describing the contributions of first order and higher order factors. Intercorrelations of selected variables can also be shown as more concrete illustrations of the effects of first and higher order factors.

Table 1 contains the distributions described above along with means and standard deviations. The first order factors in the Adkins and Lyerly data are responsible for a higher percentage of the variance than the first order factors in the Cattell data. It should also be borne in mind in interpreting these results that the available communality would ordinarily be spread over many more first-order factors than all higher order factors combined. Furthermore, several of the Cattell first-order factors are specifics so that diagonals on $R_{0.2,3...k}$ are, in a sense, inflated.

The lesser importance of Cattell's first-order factors, it should also be noted, is not critical with respect to his conclusions. His study was designed for higher order factoring. In other studies, however, in which the first-order factors are of prime importance to the investigator, the technique here being illustrated is a desirable, even necessary, check on the conclusions reached.

In both of the two sets of data the correlation between the two estimates of the diagonal is only moderately negative. This means that the effects of controlling first-order factors or higher order factors are far from homogeneous with

respect to the measures. Subsets of measures are differentially affected by factors in the several orders. Table 2, for example, contains estimated and obtained intercorrelations of selected verbal tests from Adkins and Lyerly. Obtained correlations are below the diagonal while the estimated contributions to the correlations of first-order factors alone and higher order factors alone are above the diagonal. Within each cell above the diagonal the upper value represents $R_{0.2,3..k}$, the lower value $R_{0.1}$. Thus we see that the intercorrelations of the verbal tests in this analysis tend to be explained more by higher order factors than by first-order factors.

Table 3 contains similar data for the Primary Mental Ability measures used by Cattell. First-order factors contribute only to the correlations between the parallel forms while the higher order factors account, as one would expect, for the intercorrelations of the different tests. The Fluency test for which no parallel form was available defined a specific factor in the Cattell analysis. This is clearly seen in the correlations presented. Furthermore, 39 of the possible 40 correlations with other variables in the full $R_{0.2,3..k}$ matrix are smaller than .10 for the Fluency measure, and the fortieth is less than .20.

Summary and Conclusions

A methodology has been described and illustrated for obtaining an evaluation of the importance of the factors in a particular order of factoring that does not require factoring beyond that order. For example, one can estimate the intercorrelations of the original measures with the perturbations of the first-order factors held constant or, the reverse, estimate the contribution to the intercorrelations of the original measures from the first-order factors alone. Similar operations are possible at higher orders.

An estimate of communality of the factors at a given level is required in order to estimate correlations at the lower level. When many factors are involved, squared multiples can be used for this purpose. Under these circumstances, also, the importance of the factors can be gauged by the size of the correlations of the original measures, or factors, with the reference vector structure for those measures or factors.

References

Adkins, D. C. and Lyerly, S. B. Factor analysis of reasoning tests. 1952, Chapel Hill: The University of North Carolina Press, 122 pp.

Cattell, R. B. Theory of fluid and crystallized intelligence: a critical experiment. Journal of Educational Psychology, 1963, 54, 1-22.

Horn, J. L. On subjectivity in factor analysis. Educational and Psychological Measurement, 1967, 27, 811-820.

Humphreys, L. G. The organization of human abilities. American Psychologist, 1962, 17, 475-483.

Humphreys, L. G. Critique of Cattell's "Theory of fluid and crystalized: a critical experiment." Journal of Educational Psychology, 1967, 58, 129-136.

Schmid, J. and Leiman, J. The development of hierarchical factor solutions. Psychometrika, 1957, 22, 53-61.

Footnotes

1. This research was supported by the Office of Naval Research under contracts N 00014-67-A-0305-0003, Ledyard R. Tucker, principal investigator, and N 00014-67-A-0305-0012, Lloyd G. Humphreys, principal investigator.

Table 1

Distributions of Diagonal Values in $R_{0.1}$ and $R_{0.2,3...k}$ from Two Separate Analyses

	Adkins and Lyerly (1952)		Cattell (1963)	
	$R_{0.1}$	$R_{0.2,3...k}$	$R_{0.1}$	$R_{0.2,3...k}$
65		1		
60		1		1
55		1		0
50	1	2		1
45	1	6		5
40	3	13	2	4
35	7	14	4	5
30	9	16	8	6
25	7	5	15	9
20	13	5	7	7
15	9	2	2	3
10	5		3	
05	3			
00	5			
-05	2			
-10	0			
-15	1			
$\bar{X}$	.22	.37	.28	.33
S	.13	.10	.10	.11

Table 2

Comparison of Obtained and Estimated Intercorrelations of Selected Verbal Tests from Adkins and Lyerly (1952)*

		49	50	59	60	61	62
49	Reading 1		30 32	26 32	21 38	21 36	44 29
50	Reading 2	65		21 36	21 38	23 37	27 33
59	Verbal Analogies	57	54		17 41	20 41	28 34
60	Verbal Classification 1	58	59	58		30 48	21 41
61	Verbal Classification 2	56	60	61	87		21 40
62	Vocabulary	74	65	65	63	62	

* Measures are numbered as in Adkins and Lyerly. Entries below the diagonal are observed intercorrelations; the upper one of the pair of entries above the diagonal is from $R_{0.2,3..k}$ (first order factor contributions), and the lower one is from $R_{0.1}$ (higher order factor contributions).

Table 3

Intercorrelations of PMA Measures

from Cattell (1963)*

	1	2	3	4	5	6	7	8	9
1 Verbal 1		48 38	04 27	05 26	00 43	02 39	-02 36	-02 34	08 36
2 Verbal 2	86		00 29	00 28	00 43	02 40	00 39	-02 36	09 37
3 Space 1	30	30		53 26	-02 24	00 24	-04 19	01 13	05 20
4 Space 2	32	27	79		-03 25	03 23	-04 20	01 14	05 16
5 Reasoning 1	41	42	21	23		40 38	02 41	04 41	03 34
6 Reasoning 2	42	41	25	25	77		-02 39	01 38	05 30
7 Numerical 1	34	37	23	16	40	37		44 33	04 39
8 Numerical 2	32	33	19	14	43	39	78		04 36
9 Fluency	44	45	17	22	36	33	42	40	

* Entries below the diagonal are observed intercorrelations; the upper one of the pair of entries above the diagonal is from $R_{0.2,3..k}$ (first order factor contributions), and the lower one is from $R_{0.1}$ (higher order factor contributions).

Abstract

A major conclusion of the 1947 Scottish survey of intelligence was that there had been a gain since 1932 on the group test but no gain on the individual Stanford-Binet test. This conclusion is marred, however, by the use of regression methods of equating the 1916 and 1937 editions of the individual test for which only 89 cases were available. Avoidance of the sample of 89 cases who had been administered both editions of the individual test by the use of the equipercentile method of equation reveals parallel gains for both the group and the individual test. There is no need to qualify the conclusion that a small increase in intelligence among Scottish school children occurred between 1933 and 1947.

Footnote to the Scottish Survey of Intelligence

Lloyd G. Humphreys

University of Illinois

The results of the 1947 Scottish survey of intelligence (Scottish Council for Research in Education, 1949) were somewhat ambiguous, as reported, with respect to a gain in intelligence between 1932 and 1947. The group test, which was the principal survey instrument, showed a gain. The 1916 and the 1937 editions of the Stanford-Binet were also administered in 1933 and 1947, respectively, to 500 or more students of each sex. When the two editions were equated, by a procedure which will be described briefly below, the authors reported a slight loss for girls. Overall the mean intelligence quotients were almost precisely the same. The conflict in results between the group test and the individual tests has sometimes been interpreted as meaning that "real" intelligence did not change.

The procedures used in equating the two versions of the individual test are, however, open to question. First, the authors used a regression method for equating group and individual test scores separately for the two sexes. Then they used a regression method for equating the scores on the separate individual tests. The equation of group to individual test involved 500 or more of the special cases for each sex, but the final equation of the two individual tests involved only 89 cases of both sexes combined.

A good case can be made against the regression method of equating scores on two tests each of which is supposedly measuring basically the same function. An even better case can be made for forgetting about the 89 cases who had been given both versions of the individual test and base the equation of the two on their relationship to the group test. An N of 1000 or more for this step in the procedure has clear advantages over an N of 89 but use of the large N requires the equipercentile method of conversion.

There are advantages to the use of the equipercentile method over and beyond its applicability to the data based upon the large N. It requires only ordinal scales of measurement and is thus independent of the shape of the regression. It is also independent of attentuation in the slope of the regression introduced by measurement error. It does require the assumption that the two measures are equally valid measures of the trait, but the regression method requires a different and at least equally difficult assumption that one of the two measures can be considered the criterion measure of the trait.

Table 1 contains comparable scores for boys and girls for each level of the group test on the 1916 and 1937 editions of the Stanford-Binet. Also included are the differences between the intelligence quotients for each level of the group test. When these differences are weighted by their respective Ns and averaged, it is seen that the mean difference in intelligence quotients for boys between the 1916 and 1937 editions is 3.14 units. The comparable figure for girls is -.99. There is clearly an interaction between sex and the two editions of the Stanford-Binet when the conversion is based upon the group test common to both testing periods. This same interaction is also seen in the results from the 6-day sample in which girls are slightly superior to boys on the group test and significantly inferior to boys on the 1937 revision. For some reason, Scottish girls seemed to be at a relative disadvantage on the revised Stanford-Binet in spite of the near equality of the sexes in the standardization samples. The latter were, of course, drawn entirely from the United States.

There are two ways in which the equipercentile method can be applied. Equipercentile conversions can be computed between the group test and each individual test, and then the individual tests can be converted to each other. Alternatively, the published regression conversions of individual test on group test can be

accepted, but an equipercentile conversion can be substituted for the published regression conversion between the two individual tests. The latter will be designated the "mixed" method.

Table 2 presents the published data on group test and the regression conversions of the individual tests in the first three numbered rows. These data are followed by equipercentile conversions computed by the present writer. The means in lines 2 and 4 which represent different estimates of the population I. Q.s on the individual tests are generally comparable though the equipercentile values are somewhat lower. This difference is a regression phenomenon, arising from the lack of perfect correlation between group and individual test. Conversions of 1916 I. Q.s into 1937 I. Q.s are contained in lines 3, 5, and 7. The two variations of the equipercentile method are in general agreement in showing a gain in I. Q. for girls, but both depart radically from the regression results. Furthermore, when the 1932 results are presented in the units of the 1937 revision of the Stanford-Binet (lines 6 and 8), the discrepancy between the results for the two sexes is very marked.

The weak conclusion that can be drawn from this analysis is that the previous outcome of no gain on an individual test of intelligence between 1932 and 1947 is questionable since a different and supportable methodology demonstrates a gain. The strong conclusion, which accepts the superiority of the equipercentile methodology for problems of this type, is that gains on group and individual tests are approximately parallel and that the gain was greater for girls than for boys. From the latter point of view there is no need to qualify the conclusion that a small increase in intelligence among Scottish school children occurred between 1933 and 1947. One is strongly tempted, furthermore, to predict that the gain would have been larger without the disrupting effects on education of World War II.

Even so the results are well in line with Tuddenham's demonstration of an increase in intelligence among men in the United States between World War I and World War II (1948).

Table 1

Equivalent Scores on the 1916 and 1937 Editions of the Stanford-Binet for Various Levels of the Group Test

	Boys			Girls		
Group test	1937 ed.	1916 ed.	Difference	1937 ed.	1916 ed.	Difference
69.5	167.83	149.17	18.66	154.50	149.50	5.00
64.5	151.17	140.12	11.05	149.50	139.50	10.00
59.5	134.50	128.46	6.04	133.71	132.50	1.21
54.5	125.02	121.53	3.69	122.68	119.71	2.97
49.5	116.06	112.94	3.12	113.72	112.36	1.38
44.5	108.46	106.37	2.09	107.10	107.05	.05
39.5	103.96	100.48	3.48	99.27	100.01	-.74
34.5	100.11	95.33	4.78	94.01	94.21	-.20
29.5	95.77	90.88	4.89	88.43	90.22	-1.79
24.5	89.85	87.38	2.47	82.68	86.17	-3.49
19.5	85.96	85.26	.70	79.27	82.28	-3.01
14.5	82.09	81.57	.52	75.52	79.05	-3.53
9.5	77.71	77.71	.00	72.17	75.41	-3.24
4.5	73.79	71.64	2.15	67.23	70.05	-2.82
-.5	54.50	54.50	.00	49.50	64.50	-15.00
Weighted Mean Difference			3.14			-.99

Table 2

Summary of Results for 1932 and 1947

		Boys		Girls	
		1932	1947	1932	1947
(1)	Group Test	34.503	35.880	34.409	37.622
	Regression Conversions				
(2)	Individual on Group	99.86	103.68	98.56	100.75
(3)	1916 on 1937		100.48		97.89
	Equipercentile Conversions				
(4)	Group to Individual	98.29	103.52	97.67	100.61
(5)	1937 to 1916		99.91		101.49
(6)	1916 to 1937	101.70		96.33	
	Mixed Conversions*				
(7)	1937 to 1916		100.54		101.74
(8)	1916 to 1937	103.00		97.57	

*Utilizes the regression conversion of each individual test on the group test (line 2) as the first step, but uses an equipercentile conversion as the second step.

References

Scottish Council for Research in Education, The Trend of Scottish Intelligence, 1949, University of London Press, London, 151 pp.

Tuddenham, R. D. Soldier intelligence in World Wars I and II, American Psychologist, 1948, 3, 54-56.

May 1970

OFFICE OF NAVAL RESEARCH

PERSONNEL AND TRAINING RESEARCH PROGRAMS (CODE 458)

DISTRIBUTION LIST

CONTRACT NO. N0014-67-A-0305-0012

CONTRACTOR University of Illinois
Lloyd G. Humphreys

NAVY

4 Chief of Naval Research
Code 458
Department of the Navy
Arlington, Virginia 22217

1 Director
ONR Branch Office
495 Summer Street
Boston, Massachusetts 02210

1 Director
ONR Branch Office
219 South Dearborn Street
Chicago, Illinois 60604

1 Director
ONR Branch Office
1030 East Green Street
Pasadena, California 91101

6 Director, Naval Research Laboratory
Washington, D.C. 20390
ATTN: Library, Code 2029 (ONRL)

1 Office of Naval Research
Area Office
207 West Summer Street
New York, New York 10011

1 Office of Naval Research
Area Office
1076 Mission Street
San Francisco, California 94103

6 Director
Naval Research Laboratory
Washington, D.C. 20390
ATTN: Technical Information Division

20 Defense Documentation Center
Cameron Station, Building 5
5010 Duke Street
Alexandria, Virginia 22314

1 Commanding Officer
Service School Command
U.S. Naval Training Center
San Diego, California 92133

3 Commanding Officer
Naval Personnel and Training
Research Laboratory
San Diego, California 92152

1 Commanding Officer
Naval Medical Neuropsychiatric
Research Unit
San Diego, California 92152

1 Commanding Officer
Naval Air Technical
Training Center
Jacksonville, Florida 32213

1 Dr. James J. Regan, Code 55
Naval Training Device Center
Orlando, Florida 32813

1 Technical Library
U.S. Naval Weapons Laboratory
Dahlgren, Virginia 22448

1 Research Director, Code 06
Research and Evaluation
Department
U.S. Naval Examining Center
Building 2711 - Green Bay Area
Great Lakes, Illinois 60088
ATTN: C.S. Winiewicz

1 Chairman
Behavioral Science Department
Naval Command and
Management Division
U.S. Naval Academy
Luce Hall
Annapolis, Maryland 21402

1 Chairman
Management Science Department
Naval Command and Management Div.
U.S. Naval Academy
Luce Hall
Annapolis, Maryland 21402

1 Dr. A. L. Slafkosky
Scientific Advisor (Code AX)
Commandant of the Marine Corps
Washington, D.C. 20380

1 Behavioral Sciences Department
Naval Medical Research Institute
National Naval Medical Center
Bethesda, Maryland 20014

1 Commanding Officer
Naval Medical Field
Research Laboratory
Camp Lejeune, North Carolina 28542

1 Director, Aerospace Crew
Equipment Department
Naval Air Development Center
Johnsville
Warminster, Pennsylvania 18974

1 Chief, Naval Air
Technical Training
Naval Air Station
Memphis, Tennessee 38115

1 Director, Education and
Training Sciences Department
Naval Medical Research Institute
National Naval Medical Center
Building 142
Bethesda, Maryland 20014

1 Commander, Submarine Development
Group TWO
Fleet Post Office
New York, New York 09501

1 Commander, Operational
Test and Evaluation Force
U.S. Naval Base
Norfolk, Virginia 23511

1 Office of Civilian Manpower
Management, Technical
Training Branch (Code 024)
Department of the Navy
Washington, D.C. 20390

1 Chief of Naval Operations
(Op-07TL)
Department of the Navy
Washington, D.C. 20350

1 Chief of Naval Material (MAT 031M)
Room 1323, Main Navy Building
Washington, D.C. 20360

1 Mr. George N. Graine
Naval Ship Systems Command
(SHIPS 03H)
Department of the Navy
Washington D.C. 20360

1 Chief, Bureau of
Medicine and Surgery
Code 513
Washington, D.C. 20390

1 Chief, Bureau of Medicine
and Surgery
Research Division (Code 713)
Department of the Navy
Washington, D.C. 20390

9 Technical Library (Pers-11b)
Bureau of Naval Personnel
Department of the Navy
Washington, D.C. 20370

3 Personnel Research and
Development Laboratory
Washington Navy Yard, Building 200
Washington, D.C. 20390
ATTN: Library, Room 3307

1 Commander, Naval Air Systems
Command, Navy Department
AIR-4132
Washington, D.C. 20360

1 Commandant of the Marine Corps
Headquarters, U.S. Marine Corps
Code AO1B
Washington, D.C. 20380

1 Technical Library
Naval Ship Systems Command
Main Navy Building, Room 1532
Washington, D.C. 20360

1 Mr. Philip Rochlin, Head
Technical Library Branch
Naval Ordnance Station
Indian Head, Maryland 20640

1 Library, Code 0212
Naval Postgraduate School
Monterey, California 93940

1 Technical Reference Library
Naval Medical Research Institute
National Naval Medical Center
Bethesda, Maryland 20014

1 Scientific Advisory Team (Code 71)
Staff, COMASWFORLANT
Norfolk, Virginia 23511

1 Education & Training
Developments Staff
Personnel Research &
Development Laboratory
Washington Navy Yard, Bldg. 200
Washington, D.C. 20390

1 Dr. Don H. Coombs, Co-Director
ERIC Clearinghouse
Stanford University
Palo Alto, California 94305

1 ERIC Clearinghouse on
Educational Media and Technology
Stanford University
Stanford, California 94305

1 ERIC Clearinghouse on Vocational
and Technical Education
Ohio State University
1900 Kenny Road
Columbus, Ohio 43210
ATTN: Acquisition Specialist

1 Lt. Col. F. R. Ratliff
Office of the Assistant
Secretary of Defense (M&RU)
The Pentagon, Room 3D960
Washington, D.C. 20301

1 Dr. Ralph R. Canter
Military Manpower
Research Coordinator
OASD (M&RA) MR&U
The Pentagon, Room 3D960
Washington, D.C. 20301

1 Deputy Director
Office of Civilian Manpower
Management
Department of the Navy
Washington, D.C. 20390

1 Chief, Naval Air
Reserve Training
Naval Air Station, Box 1
Glenview, Illinois 60026

1 Technical Library
Naval Training Device Center
Orlando, Florida 32813

1 U.S. Naval Submarine
Medical Center
P. O. Box 600, Naval Submarine Base
Groton, Connecticut 06340
ATTN: Dr. B. B. Weybew

1 Dr. Robert F. Lockman
Center for Naval Analyses
1401 Wilson Boulevard
Arlington, Virginia 22209