THE UNIVERSITY OF ALBERTA

ESTIMATING CONSTRUCT VALIDITY IN MULTIPLE CHOICE, ESSAY, AND SIMULATION GRADUATE ACHIEVEMENT EXAMINATIONS

BY

© CLARKE B. HAZLETT

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA

SPRING, 1972

ABSTRACT

It has been assumed that validity, in particular construct validity, is a sufficient criterion for determining the worth of achievement measures. This study was undertaken in order to investigate certain constructs (cognitive or mental behaviors) labelled as factual, comprehension, and problem solving in graduate achievement examinations in the area of applied statistics and research design. The first of these constructs--factual--approximates the behavior that Bloom calls knowledge. Comprehension approximates behaviors labelled by Bloom as comprehension, application, and analysis; and similarly problem solving approximates behaviors of synthesis and evaluation. The three constructs are assumed to have a mental hierarchical structure and are dependent upon the perceived degree of familiarity the tested information has to each individual.

The use of these three cognitive components as the analytic framework for evaluating graduate achievement examinations was considered justifiable in the light of (1) limitations found in Bloom's taxonomy, and (2) the potential use this approach offered in the general area of educational measurement and decision making.

The particular methods of testing that were investigated were (1) the multiple choice format, (2) the essay examination, and (3) the rather novel technique of computerized simulation exercises. Subsections of each method were preclassified as electing behavior which would correspond to the characteristics

of factual, comprehension, and problem solving cognitions. Since two forms of each method were administered to all tested subjects, replication of construct validity estimates was obtained.

The degree of validity each method provided was inferred by inspecting a multitrait-multimethod matrix as well as by factor analyzing the correlations in this latter matrix. Except for the simulation problem solving tests, it was generally concluded that the three traits of interest were not adequate representations of test behavior. Even the simulation problem solving estimates were of dubious value, since the tests used in establishing these estimates were regarded as the most difficult by examinees.

Caution must be used in generalizing the results of this study to all graduate testing in the area of applied statistics and research design because (1) this study failed to establish respectable estimates of alternate form reliability, (2) subjects were not randomly selected, and (3) tests were administered under reasonably non-stress conditions. The failure of this study to establish reasonable construct validity estimates is in keeping with many other similar efforts. Consequently, the proposition is made, that in the area of achievement measures, it may be more fruitful for the examiner to concentrate his efforts in the area of content validity.

DEDICATION

To Stella and Mom, for their love and confidence,
And to Dad, now only an inspirational memory.
These three -- always loved, never forgotten.

ACKNOWLEDGEMENTS

I am indebted to my supervisor, Dr. S. Hunka for his considerable assistance throughout my graduate training, and in particular during the time of this study. His kindness, instruction, and example have been an integral part of my most valuable educational experiences.

Appreciation is expressed to the readers of this thesis -- Drs. L. Fisher (External Examiner), S. Stinson, V. Nyberg, and in particular, T. Maguire.

Special thanks are expressed to those individuals who helped in the implementation of this study -- W. Osbaldeston, Dr. D. Burnett, V. Muller, and P. Harasym.

CONTENTS

CHAPTER

CHAPTER PAGE

APPENDICES

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

Statement and Importance of the Problem

In order to make educational decisions most educators still rely to some degree on information gathered from achievement tests (Dyer, 1965; Manning, 1969; Friedenberg, 1969). As Stanley (1971, p. 357) points out, some of the literature in the area of achievement measures maintains that the degree of reliance placed in such tests should be proportional to the degree of dependability and accuracy these tests exemplify. That is to say, no promotions, selections and/or placements should be based upon data that are neither reliable nor valid.

Many graduate schools (regardless of the area of specialization) base their initial selection of candidates, their course grades, and their awarding of degrees on achievement measures. Typically, multiple choice items and essay questions are used as means for collecting these data. If these data are to carry weight in the educational process, it is important the data be analyzed in order to determine if they meet the criteria of adequate consistency and validity. This study was undertaken to help determine the extent to which two traditional and one newer method of measuring graduate student achievement met these criteria (provided criterial

performances were those cognitive skills to be described later (cf. 6-14)).

While studies in this area are not new, the present investigation was not limited to the use of essay and multiple choice examinations. Simulation has become increasingly used in training and evaluation programs (Holland, 1967; Hubbard, Lenit, Schumaker, & Schnabel, 1965; Lumsdaine, 1968; MacDonald, 1965; Stolurow, 1968), particularly since the advent of the computer. Since little work has been done in establishing validity estimates using simulation, it was felt that this potential evaluative technique should also be investigated. As will be seen in the review of related literature in Chapter Two, evaluators seldom agree on the use of essays and multiple choice items. The cited advantages and disadvantages of each method create dilemmas for which simulation (at least in terms of face value) appeared to offer a compromise. For example, estimation of strategic and organizational ability is often cited to be a behavior manifested in essay compositions; this too appears to be required in most simulation exercises. Reliability of scoring, the obvious shortcoming of essays, is cited as the main advantage of multiple choice tests; if simulation also requires a selection of alternatives, it too should be consistent.

Since the underlying concept to both reliability and validity is consistency (Campbell and Fiske, 1959), there must be some defined variable(s) upon which consistency estimates can be made. This author did not try to estimate achievement test reliability and validity in terms of educational decision making. Rather the

approach of Loevinger (1957) was taken. She states "...the most fruitful direction for the development of psychometric devices, and hence psychometric theory, is toward measurement of traits... [p. 640]." Though there are exceptions (e.g., Bechtoldt, 1959) much of the literature (APA, AERA, and NCME, 1966; Campbell, 1960; Cronbach and Meehl, 1955; Cronbach, 1960; Jenkins and Lykken, 1957; Jessor and Hammond, 1957) favors the same approach as Loevinger's. Specifically Marschak (1954) pointed out the advantage of trait-oriented studies over decision-oriented studies: "Theory provides us with solutions which are potentially useful for a large class of decisions [p. 214]." Since no one can foresee all future decisions, a study aimed at the improvement of isolated selections might not change properties of the evaluation procedure that would require later modifications.

In accepting the above position, the question for this study became one of deciding which "traits" would provide a basis for estimating various types of test consistency, and which in the long run would not only help educational decision making but also contribute to the theory of measurement in general. This author looked to the medical education area for selection of these traits.

Basis for Defining Traits

In Canada, one of the largest graduate degree awarding institutions is the Royal College of Physicians and Surgeons of

Canada. The R. S. McLaughlin Examination and Research Centre is a body set up by the Royal College to investigate ways and means to assess and, if possible, improve present evaluation procedures in this accreditation process. One of the major decisions taken by the McLaughlin Centre was to define traits (or constructs) to be measured in specialty examinations. In the minutes of the first meeting of the Test Committee for Internal Medicine, May 3, 1968, a decision (which subsequently was adopted by other specialties) was recorded as follows.

> Item D Taxonomy
> After examining the proposed six point taxonomy system it was agreed to simplify it as set out below.
> Taxomony 1 - Recall, factual knowledge
> 2 - Comprehension, application, and analysis
> 3 - Problem Solving

The "proposed six point taxonomy system" referred to above was the hierarchical cognitive structure hypothesized by Bloom, et al (1956). In reducing it to three levels, the Committee had hoped to maximize the usefulness of Bloom's taxonomy for the medical education area. Firstly, by collapsing certain categories of Bloom's taxonomy it was hoped the resulting three traits would appear more mutually exlusive to the user than those in Bloom's hierarchy; if such face validity was realized, novice test writers might be more inclined to follow the guidelines of this modified taxonomy. Secondly, it was hoped that items

measuring these three traits would elicit behavior that could be more clearly delineated even though it was still assumed that behavioral patterns up through these three levels were theoretically dependent since a hierarchical structure was hypothesized. Thirdly, it was hoped the three traits proposed still incorporated two of the most often cited advantages of Bloom's classification, namely--(1) a guideline for test writers to identify a variety of appropriate behaviors to be used as criteria in evaluation, and (2) a help for shifting (medical) evaluations' exclusive emphasis on content to include behavioral operations expected in candidates' performances.

The utility and existence of the above three traits (in terms of their reliability and validity) was used as the basis of this study. The following steps were taken to provide, if possible, some generalizability to the area from which the three traits had been originally hypothesized: (1) the behavioral characteristics of these traits (as seen by the medical committee) (cf. pp. 6-14) were used in the development and selection of examinations in the area of research design; (2) only graduate students were used as test subjects.

Since the literature (Gagne, 1965; Sullivan, 1969) has criticized Bloom's classification by noting there is a lack of generalizability of the imputed mental processes across subject matter, it was considered useful to determine if these traits were present in graduate research evaluation. If later studies indicated that these traits were also present in medical evaluations, then

one limitation of Bloom's scheme might be overcome. Finally, if the condensed classification was found to be easier to follow as well as eliciting test writers skills which reflected the scheme, it could be recommended as an alternative to Bloom's classification which incorporates some definite limitations in this area (cf. pp. 28-29). Consequently, the study undertaken and reported here not only appeared interesting but necessary.

Since this study had chosen to use traits which are a modification of Bloom's taxonomy, the reader is now referred to the following section in this chapter for a more precise explanation of how these traits relate to that latter classification scheme.

Relationship to Bloom's Taxonomy

The three taxonomic levels or traits used by the Royal College have been referred to as "Factual", "Comprehension", and "Problem Solving" (Hazlett, 1969, 1970).

The factual level corresponds approximately to the same level that Bloom, et al (1956, pp. 62-88) describes as knowledge. There are at least three similarities between "factual" and "knowledge" behavior. Firstly, both are characterized by inferences not observations; that is to say, mental behavior of this kind must be inferred rather than observed directly. (This similarity is characteristic of all behavioral levels in both Bloom's taxonomy and the ones used in this study.) Secondly, both factual and knowledge behavior can be described by terms such as simple recall

or simple recognition. Thirdly, both classifications regard this behavior as characteristic of the simplest or lowest mental behavior usually measured in achievement tests. Other authors (Gagne, 1963, p. 63; Purves, 1971, p. 707; Baldwin, 1971, p. 872, Valette, 1971, p. 823; Moore and Kennedy, 1971, p. 412) similarly refer to this behavior as minimal mental skill. However, the Committee which originally proposed the three traits (cf. p. 4) made a subtle, but meaningful, distinction between facts and knowledge. Bloom, Hastings, and Madaus (1971) specify that evaluation questions measuring knowledge "...should never call for finer discrimination or more exact usage than that for which...teaching might account. If they do...[they are] testing for some behavior beyond knowledge [p. 145]." The Committee of the Royal College decided it could not presume to know the manner in which topics were taught since candidates were being evaluated from a variety of institutions. It did presume, however, that any candidate who had taken at least four years of graduate training (a requirement for sitting in accreditation examinations) should be well acquainted with certain information. To use the vernacular, such information would be classified as "finger-tip" data. Consequently, test items which closely correspond to not only what is taught, but also to what is gained through experience, were regarded in this study as factual. Information that is not only specifically taught, but well learned or used a great deal (by most if not all subjects) was considered to be elicited by minimal mental effort.

Consequently, "factual" refers to a mental behavior that is no more than an excitation of a well-used memory path (Hilgard, 1957, p. 288). Such a definition does not include any assumptions about the manner in which a path becomes entrenched. The assumption is made, however, that individuals trained within given content areas, will require minimal mental effort when evaluated on information in which they are well versed.

The second label, "comprehension", was used in this study in a much broader sense than used by Bloom et al (1956, pp. 89-119). As indicated in the previous section of this chapter (cf. p. 4) the medical personnel chose this label to describe behavior that Bloom characterized as comprehension, and application and analysis (Bloom et al 1956, pp. 89-161). However, it was recognized that, (1) some behavior that Bloom would describe as comprehension was now being classified as factual by the Royal College, and (2) application and analysis in Bloom's scheme was for the most part actually characteristics of comprehension. The modification was purposefully implemented, however, in an attempt to make descriptions of each respective mental behavior more mutually exclusive. While Bloom's descriptions of each behavioral level and sublevels are recognized as comprehensive, some writers (Gagne, 1965; Sullivan, 1969) have shown that certain categories are often distinct only in terms of their content and not necessarily in terms of any formal characteristics which affect the learning of such behaviors. In a recent attempt to refute this latter criticism, as it applied to evaluation, Bloom, Hastings and Madaus (1971) maintain levels can be made

distinguishable by the degree of familiarity the candidate has with information tested. However, their attempt at clarification is not very successful:

> ...items built for the purpose of evaluating knowledge outcomes should stick as nearly as possible to the form and level of precision used in instruction. In comprehension items...the opposite is true...[p. 150].

In other words, comprehension items should avoid eliciting information that is familiar because of prior acquaintance with it during instruction. (The reader should carefully note the context in which the word "opposite" is used.) However, in the same explanation on comprehension these authors continue:

> ...items which call for interpretation in a totally different setting from the one in which instruction was given probably come much nearer to testing application and analyses than to measuring interpretation ...[p. 150].

While it is true this reference was published after the Royal College's decision was made to collapse Bloom's categories of comprehension, application, and analysis, the reader will no doubt concur that such an explanation as given above by Bloom and his colleagues would hardly be reason for reversing its decision. That is to say, if familiarity of information (due to instruction) is the criterion for determining how mental behavior will be demonstrated then (1) independent evaluators cannot be expected to know, and (2) examinees (who have been taught by different instructors) will not uniformly demonstrate, the fine degrees of familiarity which distinguish levels in Bloom's classification. However, it was

assumed by the Committee that uniformity might be enhanced if information tested by items were classified as generally familiar and generally not familiar. In essence this dichotomy reflects the differences as well as a partial definition of the terms "factual" and "comprehension".

Consequently, the term "comprehension" may refer to such behavior as "...to comprehend the significance of particular words..." (Moore and Kennedy, 1971, p. 412), "...to extrapolate from data presented in a table..." (Orlandi, 1971, p. 476), "... to estimate or predict consequences..." (Bloom, et al, 1956, p. 96), "...to predict the probable effect of a change in a factor..." (Bloom, et al, 1956, p. 124), "...to recognize the limits within which a scientific principle is applicable..." (All India Council, 1958, p. 105), "...to apply the facts, concepts and theories to actual life situation..." (University Grants Commission, 1961, p. 123), "...to distinguish cause and effect relationships from other sequential relationships..." (Ayers, 1966), "...to identify unstated assumptions which are necessary to a line of argument..." (French, 1957, p. 179), and "...to recognize basic terms and their interrelationship..." (Tyler, 1954, p. 29). As the reader will infer from the above quotations, the term "comprehension", as used in this study can refer to mental skills which some authors have chosen to delineate and describe in a variety of ways, but all of which has one assumed factor in common: the mental reactions to the subject matter is not characterized in most subjects as being repetitive; and consequently the subject

matter is not familiar. If any of the above descriptions were common place to a candidate, however, such mental behavior would be classified as factual. Consequently, the evaluator trying to elicit either factual or comprehension behavior must know whether or not the information is familiar to most examinees.

The last category of behavior used in this study is that labelled "problem solving". This term again approximately corresponds to behaviors described by Bloom et al, (1956, pp. 162-200), in particular--synthesis and evaluation. The logic and assumed justification for collapsing some of Bloom's categories into one called "comprehension" was the same for collapsing synthesis and evaluation into this last category of "problem solving". However, similar to the mental behavior classified as taxonomy level 2 (i.e., comprehension), mental behavior called problem solving also dealt with non-familiar data. If a reasonable definition for distinguishing factual and comprehension was in terms of familiarity (i.e., familiar or non-familiar), what additional criteria are needed to distinguish comprehension and problem solving, particularly if the resulting classification scheme is to make categories less interdependent?

Firstly, it is reasonable to state that the intention was to dichotomize non-familiar information into novel and very novel categories. Since one of the objectives was to reduce Bloom's behavioral classification into a more simple form while

still retaining a hierarchical definition of mental behavior, a trichotomy of familiar, novel, and very novel was deemed valid. Secondly, in an attempt to extend the differences between comprehension and problem solving some characteristics of synthesis and evaluation were used as descriptors of problem solving behavior. Essentially behavior which demonstrated organizational competence, use of internal and external criteria, judgement, and versatility was considered to be problem solving. For example, to solve a problem, mental behavior is assumed to take place which is characterized by devising ("organizing") a plan of attack, the sequence of which is based on the worth and type of information ("use of criteria") each step provides. If certain steps are found wanting ("judgement") modification ("versatility") is made to the original plan. A series of mental activities is assumed to take place until an optimal path is found which leads the individual to the problem's solution. While it can be argued that the ability to write a composition is characteristic of this behavior, ability to compose is <u>not</u> considered to be a necessary or sufficient skill for describing problem solving behavior. Consequently some of the following phrases can be used to describe behavior which in this study is called problem solving: "...to propose ways of testing hypotheses..." (Bloom, 1956, p. 170), "...to devise...suitable experiments for testing hypotheses; to provide controls for experimental variables..." (Tyler, 1954, p. 39), "...to make logical experiments..." (French, 1957, p. 98), "...to check the validity of...inferences..."

(Tyler, 1954, p. 48), "...to determine if the data...supports the conclusion..." (Tyler, 1954, p. 53), and "...to detect fallacies in mathematical arguments..." (University Grants Commission, 1961, p. 254). (For a more comprehensive survey of relevant literature see Bloom, et al, 1971, pp. 191-223.)

Before leaving this section which has dealt with the relationship between traits used in this study and those described in Bloom's et al (1956) first work, the reader should note that Bloom's most recent treatment of his cognitive classification (Bloom, et al, 1971, Chp. 7-9) has by implication agreed with the structure proposed here. In this latter work the authors have chosen to treat knowledge and comprehension, application and analysis, synthesis and evaluation in these respective pairs. This is not to suggest that Bloom and his colleagues maintain that the elements of the respective pairs cannot be delineated from each other. Indeed this author would maintain the desirability of their specificity so that readers are made aware of behavioral varieties. Their treatment does imply, however, that the traits in each pair are very similar. Is it not possible that present day measurement may be sensitive only to behavioral levels represented by each of the pairs, rather than to each of the elements in those pairs? This study was undertaken to shed light upon such a possibility.

In summary, the operational steps one takes in developing a test item which measures one of these three traits are: (1) follow the guidelines and note the behavioral varieties and examples provided by Bloom, et al (1956); (2) determine the degree of familiarity most examinees will have with the information tested, remembering that examinees' prior experiences and/or exposures to instruction may cause the information to appear very familiar; (3) classify the item into one of three categories based on the item's assumed degree of familiarity. The possibility still exists that an item may be familiar to some subjects and novel to others. It is necessary therefore, for the test writer to classify an item in terms of its average degree of familiarity amongst all tested subjects.

Use of Traits

By using (1) a trichotomy of familiarity, (2) Bloom's classification of cognitive behavior, and (3) illustrative phrases from relevant literature, this author has attempted to delineate the characteristics of three traits called factual, comprehension, and problem solving. These traits were regarded in this study as the basis for defining criterial performances for estimating the degree of validity and reliability in multiple choice, essay, and simulation tests. The use of these inferred behaviors has (in essence) declared that this author had

accepted the development of such traits as his educational objectives in graduate achievement measurements. Since a statement of educational objectives is a useful (and for this study necessary) step in the process of estimating validity (cf. Chp. 2, pp. 26-30), the preceeding section is cited as a declaration of such statements. These educational objectives are regarded as reasonable and desirable because they lend themselves to a classification of inferable mental behavior which may provide "...us with solutions which are potentially useful for a large class of decisions [Marschak, 1954, p. 214.]." For example, in recent years the accumulated amount of knowledge in one area has been expanding so fast that no one individual can possibly learn all facts. Accordingly, it is reasonable to assume that more generalized information is obtained if emphasis in measurement is in the area of <u>using</u> content, rather than knowing it. That is to say, if the volume of content is so great to prohibit its total learning by an individual, it is more reasonable to measure that individual's ability to cope with familiar and unfamiliar data. If he indeed discriminates himself by his adequate use of novel information, one might expect that such an individual would similarly discriminate himself in the use of future expansions in that content area.

The use of certain modes of measurement (namely the multiple choice and essay formats) have been incorporated into this study because of their common use in educational measurement. Some

authors cited in Chapter Two feel that these formats do not always lend themselves to a measurement of all behavior cited as this study's objectives. As will become more apparent in the description of the simulation tests in Chapter Three, this newer testing format lends itself to the measurement of operations rather than content, particularly owing to the fact that simulation is by definition an attempt to approximate real life activity. Since complexities are inherent in real life situations, it is assumed individuals are required to display inferable mental behavior which has been described as factual, comprehension, and problem solving. Since an examination using simulated real life conditions is assumed to elicit these same traits, simulation was incorporated into this study in order to determine if it measured facts, comprehension, and problem solving any differently than either the essay or multiple choice tests.

The measurement of factual, comprehension, and problem solving traits by each method--essay, multiple choice, and simulation--was done simultaneously in this study. All behavioral reactions to any one method were assumed to be representative of all trait levels. As will be shown in Chapter Three, this author and two specialists in the area of applied statistics and research design classified certain questions in each method as eliciting one of these three behavioral levels. For example, the total amount of problem solving behavior any one candidate displayed was the

sum of the weighted questions classified as eliciting problem solving behavior. Similarily, the sums of other weighted execution points in the same method were the assigned scores for comprehension and factual behaviors.

The study was designed so that an alternate form of each method was administered on the second day of testing. Consequently the design incorporated a replication study. By correlating the same method, estimates of reliability were obtained. The correlation between same trait scores (e.g., two factual scores) across different methods (e.g., essay and simulation) established estimates of validity. The generalizability of the results of this study has some limitations, however, the reasons for which are given below.

Limitations of the Study

Owing to the fact that this study had purposefully limited its testing to the graduate level (in applied statistics and research design) the accessible population was of limited size. Testing outside this graduate institution was rejected because of difficulties and cost of interchanging the simulation examinations (which were programmed onto the IBM 1500 Computer Assisted Instruction (CAI) system.) Furthermore, the testing was done during the summer session of the University of Alberta, a busy time for graduates registered in courses. Many others were

working on their own thesis projects or were away working between winter sessions. After some difficulty, a total of 73 subjects were contacted, 50 of whom offered to cooperate as subjects. Therefore subjects were not randomly selected and certain historical effects (Campbell and Stanley, 1963, p. 6) were possible.

Furthermore, many of the subjects who did cooperate were colleagues of this author. Consequently it was possible that some subjects regarded the tests tasks mostly as (1) a measure of this author's ability in the content area, and (2) a threat since this author would be analyzing their behavior.

The two simulation examinations were possibly novel exercises; consequently the first exposure to simulation may have been a different task than that required by its alternate form. Stanley (1971, p. 405) notes that this factor alone may greatly interfere with procedures aimed at achieving parallel or alternate forms.

Finally, a possible weakness peculiar to this study was the degree of motivation this author attempted to instill in his subjects. The study was purposefully planned to <u>avoid</u> normal test anxiety, since it was assumed this phenomenon would interfere with the measurement of the behavioral traits. It is possible, therefore, that some subjects did not seriously perform in some test tasks, or at least not as seriously as if they were being tested under normal test conditions. This was particularly

possible in the essay tasks which required construction of, rather than selection of answers.

Summary

Owing to the fact that achievement measures must display a reasonable degree of accuracy and consistency before they can be used with confidence, this study was undertaken to estimate the degree of reliability and validity for some traditional and newer testing techniques, namely the essay, the multiple choice, and the simulation methods. Criterial performances were defined to be a modified hierarchy of Bloom's et al (1956) classification of cognitive behavior, namely factual, comprehension, and problem solving behavior. Since it was assumed that more generalized information can be obtained from a measurement of operational behavior rather than a measurement of content, the criterial characteristics of these traits were regarded as reasonable, desirable educational objectives for graduate achievement measurements.

The measurement of these traits was done simultaneously and because two forms of each method were administered a replication study was realized. However, limitations of the study have been reported to indicate the degree of generalizability this study may have. The most serious limitation was the lack of randomness in sample selection; historical, novelty, and motivational effects

were also possible sources of invalidity in the collected data. The steps that were taken to enhance the worth of this study, namely design implementation, warrants a separate chapter (cf. Chp. 3). In Chapter Four the analysis and interpretation of the collected data are discussed, and the conclusions and recommendations of these findings are reported in Chapter Five. Before covering these topics the reader is referred to a review of related literature (Chapter Two)--the literature used by this author as a guide to the topics and procedures studied in this investigation.

CHAPTER TWO

SOME PERTINENT LITERATURE

Introduction

The literature cited in this chapter falls into two main categories. In the first section an attempt is made to synthesize some related research in the area of evaluation, in particular--the controversy related to the advantages and disadvantages in using multiple choice and essay tests. It will be seen that the essence of the controversy is a question of deciding which one of these two evaluation techniques provides the educator with more valid information.

The relative newness of simulation as an evaluation technique for academic achievement probably accounts for the fact that little published material is available citing its advantages or disadvantages. Like the essay and multiple choice formats, however, the primary criterion for using simulation must also be validity. It is noteworthy that some of the work done relating to simulation validity (Hubbard, 1963; Levine, McGuire, and Nattress, 1970; McGuire and Babbott, 1967) has dealt mostly with its content or face, not construct, validity. To answer questions related to construct validity it is necessary to know what one is intending to measure. The specification of this intention is of course, a statement of education objectives. In section two of this chapter some pertinent writings are cited which specify the criteria these objectives must meet it they are to be of any use in the educational and evaluative process.

Validity and the Use of Multiple Choice and Essay Items

The relative worth of essay and multiple choice items for evaluation purposes has long been debated. Proponents of the essay maintain it can measure some aspects of performance not measured by multiple choice items. Stalnaker (1951), Huddleston (1954), Andrews (1968), and others indicate that preference for the essay lies in its apparent face validity. That is, the essay presents a work-sample type of problem, requiring summation and organization of what the examinee recalls as relevant knowledge. The examiner can then estimate the strategic ability of the examinee to use that knowledge in order to logically arrive at an answer or conclusion.

However, some researchers (Cast, 1939; Cast, 1940; Hazlett, Maguire, and Wilson, 1969) have shown that interjudge and intra-judge reliability of essay evaluation is so inadequate that there can remain little validity in the measure of any performance by essays. Other authors (Stalnaker, 1951; Levine and McGuire, 1970) have indicated further limitations of the essay: (1) it is inherently unreliable due to inadequate sampling, (2) invalid, because it is not designed to evaluate a number of important areas of competence, and (3) it is so unstructured and unstandardized that areas evaluated cannot be precisely measured.

On the other hand Levine and McGuire (1970) specify how multiple choice exams can show advantages over essays: (1) machine scoring, available with multiple choice items, can minimize reader inconsistency; (2) properly constructed and developed multiple choice examinations can show respectable internal consistency; and (3) a wide range of information can be sampled.

It has been said (Brogden, 1946; Loevinger, 1947, 1957; Stanley, 1971) that a test developer should not consider reliability as a sufficient con-

dition for adequate evaluation; rather it is a test's validity--content, construct, concurrent and predictive validity--that must occupy the test developer's attention. However, reliability puts a ceiling on validity in the sense that a test's true score cannot correlate with another measure's true score more than it does with itself (Gulliksen, 1950, pp. 95-98; Lord and Novick, 1968, p. 69). Therefore adequate reliability--that is adequate reader consistency, and/or adequate consistency over time and forms--must be considered a necessary condition for proper evaluation. It is also true, however, that the mere improvement of a test's reliability will not assure that the test's validity has also been affected and furthermore, the improvement of validity may even raise the reliability of the measuring instrument. Therefore, the argument between users of essays and multiple choice items has been essentially one of validity.

Since users of essays feel the subjective format has inherent validity a great many research studies have tried to find a means to improve the consistency of essay marks. From the earliest attempts--Darsie (1922), Ashbough (1924), Hulton (1925), Eels (1930), Cason (1931), Hartog (1935), Stalnaker (1937), Weidemann (1941), Winetrout (1941)--to those of more recent years--Vernon and Millican (1954), Edwards (1956), Wiseman (1956), Grant and Caplan (1957), Pidgeon and Yates (1957), Wiseman and Wrigley (1958), Fastier (1959), Nyberg (1966), Coffman and Kurfman (1968), Allard (1970)--there has been little success in finding the means for solution. Their suggestions--(1) using keys, (2) training markers, (3) specifying explicit behavioral objectives, (4) requiring short answers, (5) using various scoring and scaling techniques, (6) specifying adequate instructions to examinees--

can be characterized by Huddleston's (1954) statement: "...there is no convincing evidence in the literature that these hopes have come to fruition."

Attempts also have been made to improve the use of multiple choice items, in particular, to have items measure skills other than recognition. Bloom et al (1956), Bloom, Hastings, Madaus (1971), Hubbard (1961), Krathwohl, Bloom, and Masia (1964) have emphasized a taxonomy of objectives and have provided illustrative items of various taxonomic levels. However, Wolf (1967) and Sullivan (1969) have criticized this type of work because of the lack of precision in indicating either specific overt behaviors to be performed by the learner or the conditions under which such behaviors should be performed. Tyler (1950, 1969), Mager (1962), Gagné (1967) and others have tried to clarify, classify, and specify the manner in which objectives are to be formulated and have tried to specify the characteristics of objectives once they have been formulated. Those who push for behavioral objectives generally support objective examinations and work such as Paterson (1926), McCallough and Flanagan (1939), Marshall (1956), Kropp and Stoker (1966) and Bialek (1967) indicate multiple choice items can measure more than simple recall when specific behavioral objectives are kept in mind.

What usually happens in actual practice, however, seems to be another matter. Scannell and Stellwagon (1960), Hoffman (1962), Lawrence (1963), Miller, McGuire, and Larsen (1965), and Davis and Tinsley (1967) contend that many, if not most, multiple choice examinations measure only the recall of isolated bits of information which have little relevance to any meaningful behavioral objectives.

Medical educators have also found their task of adequate evaluation frustrating and difficult. The literature reveals that many of the already mentioned advantages and disadvantages in subjective and objective examinations

are also present in medical education. Blumer (1919) and Bridge (1956) both enumerate the standard criticisms against the essay even though more than three decades elapsed between their writings. Scott and Burke (1957) and Brooks (1961) point out limitations of the multiple choice test that are not any different than those criticisms found in other educational areas. Goldstein (1958) recommends a scoring system based on a dichotomy but Hazlett, Maguire, and Wilson (1969) show that such a system is still substandard. Cowles' (1954) suggestions for improving specialty examinations are certainly borrowed from other educational bodies, however, Moore (1954) and Hubbard and Clemans (1961) do give some concrete examples for breaking essays down into a series of objective examinations. Cowles and Hubbard (1952), Lennox (1957), Bull (1959), and Hubbard and Clemans (1960) maintain objective examinations are more valid than essays. It is noteworthy, however, that the validation procedures used in these latter articles are weak since correlations were done with questionable criteria (e.g., school grades) and no investigations were done to determine discriminant validity (cf. Chp. 3). The controversy of determining which format--essay or multiple choice--is best for medical evaluation is probably best illustrated by the following quotations. Brooks (1961) states "The examiner, in reading an essay question, finds many bases for estimating the scope and level of the student's accomplishment...an opportunity to see these things is more important than method or uniformity of grading [p. 91]". Karsner (1961) who even worked with Brooks in the National Board of Medical Examiners has a different opinion: "I favor the multiple-choice type of examination principally because it gives a fair comparison between the abilities of the candidates [p. 93]."

The dilemma facing evaluators in deciding to use either multiple choice or essay items is, as stated before, a problem related to validity. Since validity can be broadly defined as the accurate measurement of that which is intended, it is fundamentally important for the evaluator to know what he wants to measure. This information is to be found in his statement of objectives. The reader is now referred to the following review of literature related to the development, criteria, and use of such objectives.

Statement of Educational Objectives

The use and development of educational objectives can be classified into two main categories--nonbehavioral and behavioral.

Nonbehavioral objectives are often characterized by terms such as "goals", "aims", and "purpose". The National Education Association (1961) states an objective of education should be the "...development of the ability to think..."; similarly Dewey (1915) states "...education is a social service..."; Bruner (1960) views the goal of education as "...optimal intellectual development...". These statements are characterized by vague, personalistic use of language. While few would quarrel with the desirability of the intended outcomes, many would differ in their descriptions of those outcomes. The lack of specificity in describing outcomes was and is, therefore, a definite shortcoming of objectives stated in a nonbehavioral fashion.

With the advent of programmed instruction (Green, 1962; Skinner, 1958; Stolurow, 1961; Taber, Glaser, and Schaefer, 1965), military training programs (Gagné and Bolles, 1959), and the identification of the need for valid measurement in educational achievement (Dressel, 1954, 1961; Krathwohl, Bloom, and Masia, 1964; Popham, 1969) the lack of specificity in nonbehavioral objectives became more apparent. Due to financial and personnel costs involved in the implementation of military programs and programmed instruction, as

well as the requirement for criteria in order to assess validity in evaluation procedures, specific and unambiguous objectives were and are being demanded. The impetus from these three fields resulted in many authors enumerating ways and means in which objectives should be stated. Gagné (1962, 1965) specified that objectives should be characterized by completeness, lack of ambiquity, and internal and external consistenty. Romey (1968) required that observable behavior be specified, and that it not only be observable but terminable as well; others (Ammons, 1964, 1969; Anderson, 1965; Bloom, Hastings, and Madaus, 1971) maintain objectives should describe direction (not termination) of behavior, and consider inferred behavior to be relevant (not just observable). Tyler (1950), French (1957), Gronlund (1970), Kearney (1953) maintain objectives must be worded in terms of the pupil and that they must be exact and explicit about the behavior regarded as desirable. Mager (1962) states that objectives should (1) specify the kind of behavior which will be accepted as evidence that achievement has taken place, (2) specify important conditions under which behavior can be expected to occur, and (3) specify how well the learner must perform in order for the behavior to be classified as acceptable.

Stating behavioral objectives in terms of the above criteria has benefit for all aspects of education. With regard to evaluation the research (Adams, 1967; Furst, 1958; Guilford, 1967; Michael, 1968; Taba, 1962) shows that the nature of examinations determines what is learned. On this premise, unless a test reflects stated objectives, learning will not tend to reflect the objectives. In this study the taxonomy previously described (cf. pp. 6-14) is an attempt to describe behavior which is considered desirable. That is, the description of factual, comprehension, and problem solving behavior is an

attempt to meet some of the cited requirements for desired behavioral objectives. These traits stated as educational objectives are assumed to be complete and lend themselves to consistent classification (Gagne, 1962, 1965). They meet the conditions of Ammons (1964, 1969), Anderson (1965), Bloom, et al (1971) in that directed, inferred behavior is classified as important. They meet Mager's (1962) first conditions in that criterial behavior is classified as appropriate reaction to varying degrees of familiar data. Using these traits as one's statement of behavioral objectives, it becomes possible to (1) estimate the degree to which a particular evaluation instrument measures that behavior, and (2) estimate if examinees' behavior in handling the information contained within the instrument is similar to the type of performance described in the taxonomy.

There are disadvantages and limitations, however, in trying to state and/or use behavioral objectives. Atkin (1963, 1969), Eisner (1957, 1969), MacDonald (1965), Popham (1969) cite various problems: (1) the amount and complexity of educational outcomes are impossible to classify; (2) pre-classification schemes place unreasonable constraints on the instructor and evaluator; (3) not all outcomes are amenable to measurement; (4) not all outcomes can be anticipated; (5) only innocuous goals can be delineated; and (6) it is dehumanizing to predestine behavior.

Further criticisms have been made of Bloom's classification of objectives (Gagné, 1965; Sullivan, 1969); (1) the classification does not specify conditions under which such behavior can be expected to occur; (2) the classification lacks precision in indicating specific overt behavior to be performed by the learner; (3) categories of behavior are often distinct

only in terms of their content; (4) there is a lack of evidence that there is any generalizability of the imputed mental processes across subject matter; and (5) categories are not suitable to distinct classes of behavior for which optimal learning strategies can be specified.

The above criticisms levelled against behavioral objectives in general, and Bloom's classification in particular, can be summarized thus: it is very difficult to generate precise, explicit objectives for rather vague goals. Or to put it another way, it is impossible to describe behavior as desirable, if not only the behavior but also the reasons of desirability cannot be verbalized. However, the criticisms against behavioral objectives in general can be refuted to some degree. If educational goals are too numerous to classify, they are also too numerous to teach, consequently they are not learned in school and need not be evaluated--similarly for the arguement of complexity. Explicit objectives can hardly be considered a constraint if in fact they provide a guideline for efficient, effective instruction. If certain outcomes are not amenable to measurement then those outcomes are not amenable to judgement and consequently not crucial for educational decision making. If innocuous goals are the only goals indentifiable, then at least they are recognized as such and attempts can be made to shift educational emphasis. If setting explicit, desirable goals is dehumanizing then society at large--the family, the school, the church, the government, etc.--needs to be modified as well. It is the opinion of this author that educational progress can only be realized when educators delineate their goals, expose them to inspection and criticism; if they are found wanting at least they are known to be deficient.

The specific criticisms levelled against Bloom's classification are, in the opinion of this author, more valid than those made against behavioral objectives in general. In an attempt to rectify the weaknesses of Bloom's scheme (cf. p. 25) this author chose to use a classification scheme of inferred behavior that was similar but in some ways distinctive from Bloom's (cf. pp. 6-14). By describing behavioral reactions to information that is either familiar, novel, or very novel, by testing graduate students in nonmedical areas (even though authors of the scheme had defined its structure and usefulness in terms of medical behavior patterns), and in general by attempting to make the descriptions of behavioral levels less interdependent, this study hoped to rectify some of the cited limitations of Bloom's categorizations.

Summary

The literature cited in this chapter first dealt with the controversial uses of multiple choice and essay items in evaluation. Since validity is the central problem in this controversy as well as being the criterion for using or not using simulation, a possible resolution lay in the area of educational objectives. After reviewing the development, the criteria, the advantages, and the limitations of educational objectives the opinion was expressed that a declaration of goals worded explicitly, and in a behavioral context, was not only desirable but necessary in order for one to (1) estimate the sensitivity of a measuring instrument to such behavior, and (2) estimate if information contained within any particular instrument elicited behavior that was previously described in the evaluator's objectives. That is to say, a statement of educational objectives is a prerequisite to estimating the validity of any measuring instrument.

For this study, previously defined traits (cf. pp. 6-14), were chosen as the objective in graduate achievement measurements. The manner in which these estimates were established and estimated is the next topic of discussion (cf. Chp. 3).

CHAPTER THREE

SOLUTIONS TO THE DESIGN AND ANALYSIS PROBLEMS

Introduction

The reader is now provided with the experimental design that was used in this investigation. Since the emphasis of this study is the estimation of construct validity the first two sections will acquaint the reader with statistical procedures which have been developed and/or used for that very purpose. The manner in which traits were classified, scored, and analyzed in each of the essays, multiple choice tests, and simulation programs constitutes the bulk of this chapter. Trait classification for all tests was done *a priori* by two specialists as well as this author in order that a more accurate classification might be realized. The weighted scoring schemes for each of the method-trait subtests are also explained along with an introduction to those statistical analyses that were done on each of the subtests. Owing to the fact that evaluations using simulation have not been standardized in the educational field, the reader is given an explanatory flow-chart in order that he can more fully understand the manner in which trait scores were obtained from this novel technique. This flowchart also provides the reader with a frame of reference for determining the

generalizability of this study's conclusions about simulations used as an evaluative tool. The last two sections of this chapter describe the nature of the sample, testing procedures, as well as steps taken to assess estimates of face validity.

Multitrait-Multimethod Matrix

Campbell and Fiske (1959) have developed procedures by which convergent and discriminant validity estimates can be obtained from an intercorrelation matrix of multitraits and multitmethods. These authors point out that evidence of construct validity is not sufficiently demonstrated by reporting convergent validity estimates--i.e., the correlation of independent methods assessing the same construct or trait. They show that many tests which show respectable convergent validity are actually invalid because they also correlate too highly with other tests which are purported to measure different things. When applied to this study, Campbell and Fiske's multi-trait-multimethod correlation matrix took on the form of the model in Figure 1. The three methods--essay, multiple choice and simulated program--each yielded a subscore for three traits--factual, comprehension, and problem solving. Since this study used two forms of each method, one can see in Figure 1 that an 18 x 18 correlation matrix was obtained.

In Figure 1 the three taxonomic levels are designated by numbers 1-3, 4-6, and 7-9 respectively for the first battery of instruments and 10-12, 13-15, and 16-18 for the second battery. The reader will note that the 18 x 18 matrix has been subdivided into 4 smaller matrices, designated as A, B,

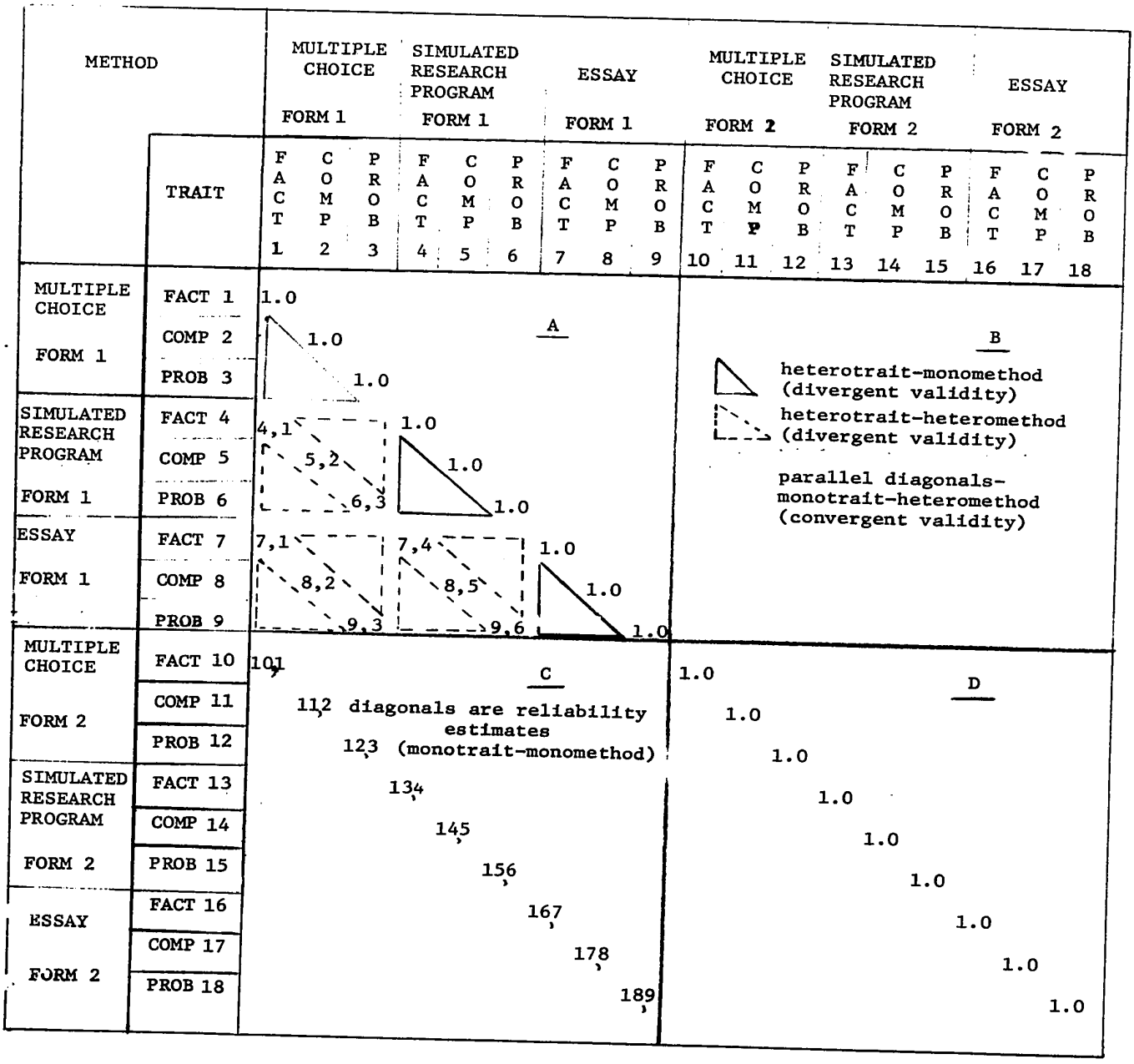


Figure 1: Multitrait-Multimethod Matrix

C, and D. A provides intercorrelations within the first battery of tests, while D contains similar estimates for the second battery. C and B contain identical intercorrelations between the two batteries.

Campbell and Fiske (1959) state "Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods [p. 83]." Reliability and validity, therefore, are actually concepts of agreement in which the interpretations applied to each vary as a function of the methods. In Figure 1 reliability estimates are contained in the diagonal elements in C (i.e., 10,1; 11,2; ...; 18,9). These estimates indicate the consistency between alternate forms measuring the same trait. Reliability estimates for this study were actually monotrait-monomethod values obtained by administering two alternate forms of the same instrument.

Convergent validity estimates (i.e., monotrait-heteromethod values) in A are specified by cells 4,1; 5,2;...; 9,6 as well as 7,1; 8,2; 9,3. In essence, these values indicate the degree of relationship between two different instruments attempting to measure the same trait. In the literature validation is typically reported only via these coefficients. However, for the validation of test interpretation and for the establishment of construct validity, convergent validity values alone are not enough. If one is to be assured that the degree of relationship indicated by monotrait-heteromethod values (i.e., convergent estimates) is influenced by the sensitivity of the measurement to a trait and not due to the influence of extraneous factor(s) (e.g., positive correlations between test-forms) then one must also look for evidence of discriminant or divergent validity.

In other words this investigator had hoped to find that different methods measuring the same trait (i.e., convergent validity values) would show a higher degree of relationship than (1) different traits measured by the same instrument, or (2) different traits measured by different instruments. Furthermore, if any relationship did exist between different traits it was hoped that the pattern of such relationships was consistent within and across methods. That is to say, whatever pattern of relationship existed between the constructs used, the pattern should have been consistent regardless of the measuring instrument if discriminant validity was present.

In this study, evidence of discriminant validity was inferred by comparing convergent validity values with (1) heterotrait-monomethod values--in A these latter values are designated by the elements in the solid triangles lying just off the diagonal elements, and (2) heterotrait-heteromethod values--in A these elements are in the dotted triangles which lie just off the convergent validity estimates. Since the traits used in this study were conceived to be a hierarchy of cognitive skills, it was hypothesized a positive correlation would exist between the traits, with factual-comprehension and comprehension-problem solving traits showing closer relationships than factual-problem solving traits. It was also hypothesized that these correlational patterns amongst traits were consistent in all heterotrait-monomethod and heterotrait-heteromethod triangles in A.

The reader will have noted that the discussion of convergent and discriminant validity for A is also applicable to corresponding elements in D. Therefore the use of the patterns in D served as a replication. In summary, to establish construct validity the multitrait-multimethod matrix should yield results similar to the following:

	Correlational Values				
	Reliability		Convergent Validity		Divergent Validity
1≥	monotrait-monomethod	≥	monotrait-heteromethod	≥	heterotrait monomethod and heterotrait heteromethod

The actual results obtained for Figure 1 are those provided in Chapter Four (cf. pp. 82-99).

Campbell and Fiske's appraisal is informal and cumbersome. Furthermore, it does not allow the researcher to clearly distinguish the degree of method covariance and trait covariance. Finally, the technique provides no explicit means for handling extraneous variance, (i.e., error variance and/or variance that is not accountable by either trait or method). Consequently a more rigorous statistical model was used to analyze the multitrait-multimethod matrix, namely that of restricted maximum likelihood factor analysis.

Restricted Maximum Likelihood Factor Analysis

The solution of a restricted maximum likelihood factor analysis has been developed (Joreskog, 1967) and programmed for computer implementation (Joreskog and Gruvaeus, 1967). This latter program has been incorporated into a more generalized computer program (Joreskog, 1970; Joreskog, Gruvaeus, and van Thillo, 1970) which, because of its availability, was the actual program used for analysis in this study.

Some recent publications (Centra, 1971; Boruch and Wolins, 1970; Boruch, Larkin, Wolins, and MacKinney, 1970; Joreskog, 1971) have illustrated the use of Joreskog's procedure for analyzing the multitrait-multimethod matrix. Using the same or similar data that Campbell and Fiske have discussed, these authors have shown that the results from Joreskog's procedure not only agree with the interpretations of Campbell and Fiske, but also provide a more succinct analysis of the data than that provided by the inspection of the multitrait-multimethod matrices.

For this study, the initial model was considered to be

$$Y_{i(jk)} = A_{jk} X_i + B_{jk} X_{ij} + C_{jk} X_{ik} + e_{i(jk)}$$

where

$Y_{i(jk)}$ = observation on the ith subject for trait j, using method k,

X_i = score of ith subject on the hypothetical general factor,

X_{ij} = score of ith subject on hypothetical factor associated with trait j,

X_{ik} = score of ith subject on hypothetical factor associated with method k,

$e_{i(jk)}$ = error associated with subject i, trait j, method k.

It was assumed that

$$Y_{i(jk)} \sim N(0,1), \qquad X_i \sim N\ I\ D(0,1),$$

$$X_{ij}, X_{ik} \sim N(0,1), \qquad e_{i(jk)} \sim N\ I\ D(0,\sigma_{e_{jk}}^2)$$

$i = 1, 2, \ldots, N$
$j = 1, 2, \ldots, J$
$k = k, 2, \ldots, K.$

The unknown parameters were

$$A_{jk},\ B_{jk},\ C_{jk},\ \sigma_{e_{jk}}^2.$$

A subject's measure was assumed to be a function of a trait (e.g., problem solving behavior), a method (e.g., essay), and a general factor (e.g., test wiseness, general achievement, etc.). The differential weights for each factor depended upon which trait-method combination came into consideration. That is to say, the general factor was weighted into the observation $Y_{i(jk)}$ according to relative magnitude of A_{jk}, the trait score (X_{ij}) and method score (X_{ik}) were weighted by B_{jk} and C_{jk} respectively. The model used was a restricted factor analytic model because certain factor loadings as well as certain correlations between factors were constrained to be zero (cf. Figures 2A, 2B).

The procedure was initiated by specifying the elements in the factor loading matrix (L) and the factor correlation (P) according to a reasonable hypothesis. The most reasonable expectation in this study was the existence of a general factor (e.g., general achievement), three method factors (corresponding to multiple choice, simulation, and essay), and three trait factors (factual, comprehension, and problem solving). Such a model is illustrated in Figure 2A. The columns of L in Figure 2A represent the general factor, three method factors, and three trait factors. The rows represent various trait-method variables (the first multiple choice tests for factual trait (M_{1F}) to the second essay for problem solving trait (E_{2P})). The reader will note that if an effect (either trait or method) was not incorporated into a measure, then there was no

	General Factor	Three Method Factors: M.C. (I)	Sim. (II)	Ess. (III)	Three Trait Factors: Fact. (1)	Comp. (2)	Prob. (3)
MC_{1F}	A_{1I}	B_{1I}	0	0	C_{1I}	0	0
MC_{1C}	A_{2I}	B_{2I}	0	0	0	C_{2I}	0
MC_{1P}	A_{3I}	B_{3I}	0	0	0	0	C_{3I}
MC_{2F}	A_{1I}	B_{1I}	0	0	C_{1I}	0	0
MC_{2C}	A_{2I}	B_{2I}	0	0	0	C_{2I}	0
MC_{2P}	A_{3I}	B_{3I}	0	0	0	0	C_{3I}
S_{1F}	A_{1II}	0	B_{1II}	0	C_{1II}	0	0
S_{1C}	A_{2II}	0	B_{2II}	0	0	C_{2II}	0
S_{1P}	A_{3II}	0	B_{3II}	0	0	0	C_{3II}
S_{2F}	A_{1II}	0	B_{1II}	0	C_{1II}	0	0
S_{2C}	A_{2II}	0	B_{2II}	0	0	C_{2II}	0
S_{2P}	A_{3II}	0	B_{3II}	0	0	0	C_{3II}
E_{1F}	A_{1III}	0	0	B_{1III}	C_{1III}	0	0
E_{1C}	A_{2III}	0	0	B_{2III}	0	C_{2III}	0
E_{1P}	A_{3III}	0	0	B_{3III}	0	0	C_{3III}
E_{2F}	A_{1III}	0	0	B_{1III}	C_{1III}	0	0
E_{2C}	A_{2III}	0	0	B_{2III}	0	C_{2III}	0
E_{2P}	A_{3III}	0	0	B_{3III}	0	0	C_{3III}

[L]

Figure 2A: Factor Loading Matrix
(Representing the Hypothesized Factor Structure)

FACTORS

FACTORS	G	I	II	III	1	2	3
G	1						
I	0	1					
II	0	$r_{II\ I}$	1				
III	0	$r_{III\ I}$	$r_{III\ II}$	1			
1	0	r_{AI}	r_{AII}	r_{AIII}	1		
2	0	r_{BI}	r_{BII}	r_{BIII}	r_{BA}	1	
3	0	r_{CI}	r_{CII}	r_{CIII}	r_{CA}	r_{CB}	1

[P]

Figure 2B: Correlation Among Factors
(Corresponding to the Factors of Figure 2A)

weighting of that effect into the measure (i.e., a value of zero was assigned). If a particular trait (or method) was used in a measure, then the factor loading corresponding to that trait (or method) was left free to vary. In summary, non-zero entries in Figure 2A were to be estimated factor loading parameters, and zero entries corresponded to parameters that were assumed to be exactly that value.

As shown by Boruch and Wolins (1970, p. 562) the inclusion of a general factor (like that hypothesized in Figure 2A) is a conservative approach relative to obtaining high loadings on trait factors. That is to say, whenever test scores are allowed to load on a general factor as well as trait factors, the proportion of test score variance accounted for by the trait factors is likely to be less than a model which does not include a general factor. Since the literature in the area of achievement measures also hypothesizes the existence of an ability such as general achievement, the model of Figure 2A appeared justified.

The hypothesized model for the correlation matrix of factors (P) is illustrated in Figure 2B. Since one of the objectives of this study was to estimate the sensitivity of various methods in measuring the hypothesized three traits, trait and method factors were allowed to correlate, but all of which were assumed to be orthogonal to the general factor. Accordingly all correlational parameters except for the first column of P were estimated, under the constraint that the first column had corresponding parameters of zero.

To determine the estimates of the unknown entries in Figure 2A and 2B the maximum likelihood method was used to fit these models to the observed data. This method minimizes the function G, where

$$G = \log|\Sigma| + tr(A\Sigma^{-1}) - \log|A| - q$$

and Σ is the population dispersion matrix, A is the observed dispersion matrix and q is the total number of variables.

The function G is a transform of the likelihood function under the assumption that the observed variables have a multivariate normal distribution. The essential formulae and basic algorithm for the minimization method are given in Joreskog (1970, Part I). The iterations of the minimization method should continue until the minimum of the function is found, "...the convergence criterion being that the magnitude of derivatives be less than 0.0005 [Joreskog, et al, 1970, p. 6]." If the minimum is reached estimators are usually accurate to the third decimal place, and approximate standard errors ($\sigma_{e_{\cdot k}}^{2}$) are defined as that proportion of variance in a measure which is not accounted for by the hypothesized factors. However, the method of maximum likelihood yields estimators which are biased, and consequently large sample sizes are needed to yield results such that the distribution of these functions in repeated samples will concentrate near the true values.

To minimize the function, however, Joreskog's (1970) computer program does not necessarily stay within theoretical constraints. Consequently, fitting a model such as that given in Figures 2A and 2B to observed data does not necessarily yield meaningful results. If such is the case, the model must be altered in order to eliminate undesirable properties of the original model. Rejection of any one model is based on several criteria.

First is the criterion of boundary limits. That is, for a particular solution all factor loadings and factor correlations must

be (within rounding error) between plus or minus one, and any specific error variance (σ_e^2) should be $o<\sigma_e^2 \leq 1$. For example, finding the correlation between two trait factors is close to one, indicates that one should collapse the appropriate columns in Figure 2 to reduce the number of factors in the altered model.

The second criterion is that of solution consistency. The correlation matrix of factors must conform to the usual restrictions on a correlation matrix; that is, P has to be positive-definite.

The third criterion is that of the chi-square test for goodness of fit. Consider the test

$$\chi 2 = (N'-1) \sum_{q<q'} (A_{qq'} - \hat{V}_{qq'})/\hat{E}_{qq}\hat{E}_{q'q}; \quad (q,q'=1,2\ldots, \text{total number of variables})$$

where $\hat{V}$ = $L \cdot P \cdot L' + \hat{E}$ = estimated correlation matrix,
A = observed correlation matrix, and
N' = constant which is a function of sample size and free parameters.

If chi-square is significant, more factors are needed assuming that the distributions of the observed variables are normal and that a linear model is appropriate. It is to be noted that smaller residuals are associated with smaller specific factors, and large residuals indicate a more inferior fit. Also, the smaller estimates of σ_e^2 are associated with larger estimates in L. To the extent that fewer factors result in higher residuals, a compromise must be reached between the criteria of number of factors and magnitude of residuals. Past practice has used a χ^2 with a corresponding probability level of 0.05 or greater for determining the optimal number of factors.

In summary the assumed advantages gained in using the factor analytic model were considerable. (1) Definition of factors would be

reasonably valid since the total correlation matrix in Figure 1 would be factor analyzed. Consequently, even reliability estimates would contribute to dimensionality of resulting factors. (2) Allowing trait factors to correlate would allow the investigator to determine the relative degree of relationships between the hypothesized constructs. (3) Including oblique method factors would allow one to estimate the degree of trait covariance and method covariance. (4) The degree of relationship observed between method and trait factors would allow the author to determine more succinctly which method was the most appropriate format for measuring each trait. (5) All variance not accountable by the hypothesized factors could be treated as error and used as a basis for determining the viability of the hypothesized model.

The inspection of Campbell and Fiske's model (and its analysis via a factor analytic technique of restricted maximum likelihood) was assumed to be a reasonable means for determining the sensitivity of essay, multiple and simulation tests in measuring cognitive behavior. The manner in which scores were obtained for these cognitive traits was, of course, crucial in a study of this kind. The following three sections deal with this explanation; the first deals with trait scores in the administered multiple choice tests, the second and third with the essays and simulations. For all three sections the descriptions provided previously (cf. pp. 6-14) were used as a basis for item development, its trait classification, and its score. In essence, therefore, these descriptions were used as this study's statement of behavioral objectives.

Development of the Multiple Choice Measures

A pool of over 500 multiple choice items were made available to this author. Using criteria such as clarity, suitability to taxonomic classifi-

cation, previous user satisfaction (from an evaluator's viewpoint), difficulty levels, and biserial correlations within particular examinations--this author chose what he considered to be the most suitable items for use in this study; a total of 85 items were so obtained. It was found, however, that items having face validity for measuring problem solving behavior were scarce. Consequently five additional problem solving items were written by this author. Furthermore, among the first 85 items some modifications were made in order that all items would have one correct choice and 4 distractors. (Consequently, all items were dichotomously scored as 1 (correct) or 0 (incorrect).) Original composition of multiple choice items was kept to a minimum in order to insure that the items that were used had previously demonstrated some characteristics of quality. (As will be shown later (cf. pp. 112 -115) the majority of these items continued to demonstrate a degree of quality if one chose to use biserial coefficients as indices of discrimination.

Having preclassified the above 90 items into factual, comprehension, and problem solving categories this author randomly assigned items from each of these three classifications into two tests, each test having 45 items. (Random assignment being defined as the alternate assignment of items from each class to each test, after the items in each class had been shuffled as a deck of cards.) Items for each test were then shuffled to obtain an assumed random order of items measuring the three traits; this order was used in the final typed copy of each test. The two tests were arbitrarily labelled test one and test two, and hereafter are referred to as such. (The numeric value of one and two did not reflect, necessarily, the order in which they were administered (cf. pp.78-79)).

These two tests were then given to two specialists (each specialist had a doctoral degree in behavioral research and applied statistics). Knowing the characteristics of the three traits they independently classified each of the 90 items as factual, comprehension, or problem solving, as well as indicating what they believed to be the correct answer for each item. Seventeen of these 90 items were not unanimously classified into one of the three categories, that is to say, one of the specialists differed with this author's classification and/or the other specialists' assignment. Final assignment to any one trait for each of these 17 items was done after unanimous agreement had been reached amongst these three individuals. Little or no differences existed in the determination of the correct answers.

After final classification, the first multiple choice test had 15 factual, 18 comprehension, and 12 problem solving items; the second multiple choice test had 14, 18, and 13 items for factual, comprehension, and problem solving respectively. These values, therefore, indicate the maximum trait score possible for each subtest. Since parallel forms have by definition equal true and error score variance as well as the same number of items (Gulliksen, 1950, p. 26), the two multiple choice tests for factual, comprehension, and problem solving were classified as <u>alternate forms</u> rather than parallel forms.

Appendix A provides the reader with the two multiple choice tests. Beside each item is a two character code indicating the assumed trait and the consecutive number of the trait within any one test. For example, a code of C3 would indicate that the item was the third comprehension item in that test.

As indicated in the directions of these two tests (cf. Appendix A) subjects simply marked their answers in the test booklet. In order to insure the accuracy of marking these booklets, two completely independent scoring procedures were used. Firstly, two different markers each scored every test and proper corrections were made whenever differences were found. Secondly, all answers were transferred to optically scored IBM answer sheets, which when scored, produced punched card data for all items. A scoring program, written in Fortran IV, analyzed and produced subtest scores for each individual. Computerized and manual scoring were then compared and any differences found were checked and corrected. In the final analysis this author was sure that 100% scoring accuracy was obtained for the multiple choice tests.

In addition to the reliability estimates reflecting consistency of multiple choice trait scores via alternate forms (Figure 1, elements 13,4; 14,5; and 15,6), inter-item consistency estimates were also calculated for each multiple choice test, for each subtest, as well as for combined subtests for each trait. The estimates were established by the Kuder-Richardson formula 20 (KR_{20}). The KR_{20} coefficient is commonly used to estimate item homogeneity, and is the average of all possible split-half coefficients for a test administered at one time. Originally developed by Kuder and Richardson (1937) and later discussed by Hoyt (1941), Gulliksen (1950), Cronbach (1951), Thorndike (1951), Tryon (1957), and Lord and Novick (1968) this coefficient is applicable when individual components are binary items which take values of zero and unity with respective probabilities of $1-p_i$ and p_i (where p_i is the proportion of candidates marking item i correctly). (The alpha coefficient is the more generalized coefficient, being applicable to continuous data.) The calculated size of

KR_{20} or alpha is a function of the (1) error of measurement, (2) the unifactoredness of the composite measure, and (3) the homogeneity of the sample to which the measure was administered. The estimated size of the coefficient indicates a test's internal consistency; if a test is composed of groups of items each measuring a different factor it then becomes difficult to know which factor to invoke to explain the meaning of a single score. Cronbach (1951), however, has shown that it is not essential that all items be factorially similar; what is required is that a large proportion of the total variance be attributable to the principal factor running through the test. In terms of the assumptions underlying this model Tryon (1957) maintains that none are required, if the coefficient is interpreted as the lower bound of reliability.

Having described the manner in which the multiple choice tests were developed, classified, scored, and analyzed the reader's attention is now directed to a similar description for trait scores in the implemented essay tests.

Development of the Essay Measures

Huddleston (1954) has indicated that most of the research done in the area of improving essay questions has shown little fruition. This study incorporated many of the suggestions the research has made (cf. p.23) however, in the hope that some benefits of these procedures might be realized if they existed. The following description will illustrate the steps taken.

After composing the essays and their keys, the author independently submitted the documents to two specialists, (both having similar qualifications to the specialists used in classifying multiple choice questions). After a series of iterative modifications the final documents are those given in Appendix B.

The reader will note that both essays have delineated the topics around which a subject's answer was to be composed. This was done since research (cf. p. 23) had indicated that all subjects should "run the same race"; that is to say, all respondents should have the same idea as to what was required in an answer. Furthermore, both keys were assumed to reflect the delineated topics in terms of factual, comprehension, and problem solving behavior. Factual marks were obtained if the subject simply used the terms specified in the two factual keys. No penalty was incorporated into the factual score if these terms were used incorrectly, though marks were lost in the comprehension section if incorrect use was evident. Accordingly, the key and the scoring scheme were assumed to reflect the behavioral characteristics previously defined (cf. p. 6-14). The problem solving score reflected the most complete answer, characterized by sophistication and judgement in the area of applied statistics and research design. For example, essay one describes a research problem which requires no inferential analysis since the entire population has been tested. Accordingly, demonstration of problem solving behavior was assumed to exist if the subject realized this and accepted the second hypothetical researcher's analysis as the correct procedure. The key for essay one describes further accepted behavior assumed to be characteristic of problem solving skill, and similarly for essay two and its key.

The written answers which the subjects provided were typed before the markers scored them. This was done to reduce bias due to legibility and any possible recognition of handwriting style by the markers. Furthermore, names of subjects were substituted for unique identification numbers, known only to this author and his two typists. Since every essay had a different identification number, markers would have had difficulty in matching any one subject's two essay answers.

Nine potential markers were qualified to score the essays, seven of whom agreed to participate. Of these 7, 2 were randomly selected on computer. (Monetary costs prohibited the use of more markers.) The two markers were doctoral candidates with previous experience in marking laboratory exercises in the area of applied statistics and research design. They were, however, well acquainted with many of the subjects. Consequently, the use of identification numbers and the typing of respondent's answers appeared to be even more justified in order to enhance any subject's anonymity.

After selecting the two markers, this author held a series of training sessions with both markers--a total time of approximately six hours. During this time the markers developed their own answers and made comparisons with the essay keys. An additional dozen practice essays were independently marked by each marker, and subsequently discussed in order that unanimity between markers would be enhanced. This author encouraged both markers to follow the key as closely as possible; slavish adherence to it was to be avoided, however, since trait scores for subjects were not to reflect "verbal dumps". That is to say, <u>length</u> of a composition was not to be

considered a sufficient, or even the most influential, criterion for discriminating among respondents. Consequently a respondent who wrote a reasonably short, succinct essay and received a high score in the problem solving subtest was not to be penalized in the factual subtest simply because his answer was incissive. This position was justifiable since the traits are regarded as an hierarchy; or put another way, demonstration of problem solving behavior requires a prerequisite ability in the factual and comprehension domains.

Furthermore, the reader will note that each key has individual scores within each trait subtest (cf. Appendix B, keys). The markers were to use these as a reference base for determining the over-all trait score, including those cases in which marks were to be given in the lower trait levels for an essay that was succint and correct in the problem solving subtest. In the latter case, a correct problem solving answer could infer some particular piece of (for example) factual knowledge. Those pieces were delineated in the factual key and were assigned specifics scores; summing such scores provided the overall factual score. Due to the time and effort involved in marking each individual "inferred" score, however, the markers often assigned only an over-all trait score. Consequently, this study could not analyze each item in the key. Thus all reported statistical analyses on the essays (cf. pp.120-123) deal exclusively with gross trait scores.

For marking purposes the essay judges were given separate xeroxed copies of each typed essay answer along with as many copies of the

respective keys as there were answers. Since each marker received separate copies of each respondent's essay, as well as marking scores on separate answer sheets (i.e., the xeroxed keys), relative independence in the assignment of marks could be assumed.

Using the analysis of variance model (Maguire and Hazlett, 1969; Winer, 1962, pp. 126-129) three different interjudge consistency estimates were established. The first of the three estimates reflected the degree of agreement for means, variances, and linear relationship between the assigned scores of the two judges; the second reflected the consistency of variance and linear relationship, the third reflected only the consistency of linear relationship. Consistency of linear relationship is simply the well known Pearson product moment correlation, and is an appropriate estimate of interjudge agreement when the scale used by each judge is identical. For this study correlation coefficients were suitable for determining interjudge reliability since the final trait mark each subject received was the average of his standardized scores assigned by each judge.

Development of the Simulation Measures

Introduction

The manner in which scores for each trait were obtained in the simulated research programs is best explained after the reader thoroughly understand the characteristics of such a testing instrument. Since this is a relatively new procedure the following subsection is devoted to this explanation.

Simulation Characteristics

Patient management problems, commonly used for purposes of evaluating medical graduates, represent one application of simulation techniques to the problems of measurement. Hubbard (1963), McGuire and Babbott (1967) have specified the criteria that patient management problems must meet in order to properly simulate physician-patient encounters. In this study their criteria were generalized in order that they would apply to the area of research design: (1) information must be in a form typically encountered in a research problem, and not in a form that is simply a summary of the problem's salient features; (2) the problem must allow a series of sequential, interdependent decisions in research management and analyses; (3) the results of each decision must be seen by the examinee in a realistic form; (4) each decision and its consequence must be binding; (5) the program must allow a variety of paths for attacking the problem, as well as providing for a variety of information (the latter being dependent upon the examinee's approach to the simulated problem). These criteria were used as guidelines in this study for the development of simulation programs in applied statistics and research design. The flowchart in Figure 3 illustrates the possible paths examinees could take in solving the simulated research problems. Both simulation programs followed the general outline of Figure 3 and consequently the following explanation applies equally to each alternate form. In general both problems required the candidate to (1) collect information regarding the actual work to be done, (2) to choose a proper measuring device to be administered, (3) to choose a proper sample, (4) to analyze the data by using statistical computer programs, and (5) to make a report of his findings.

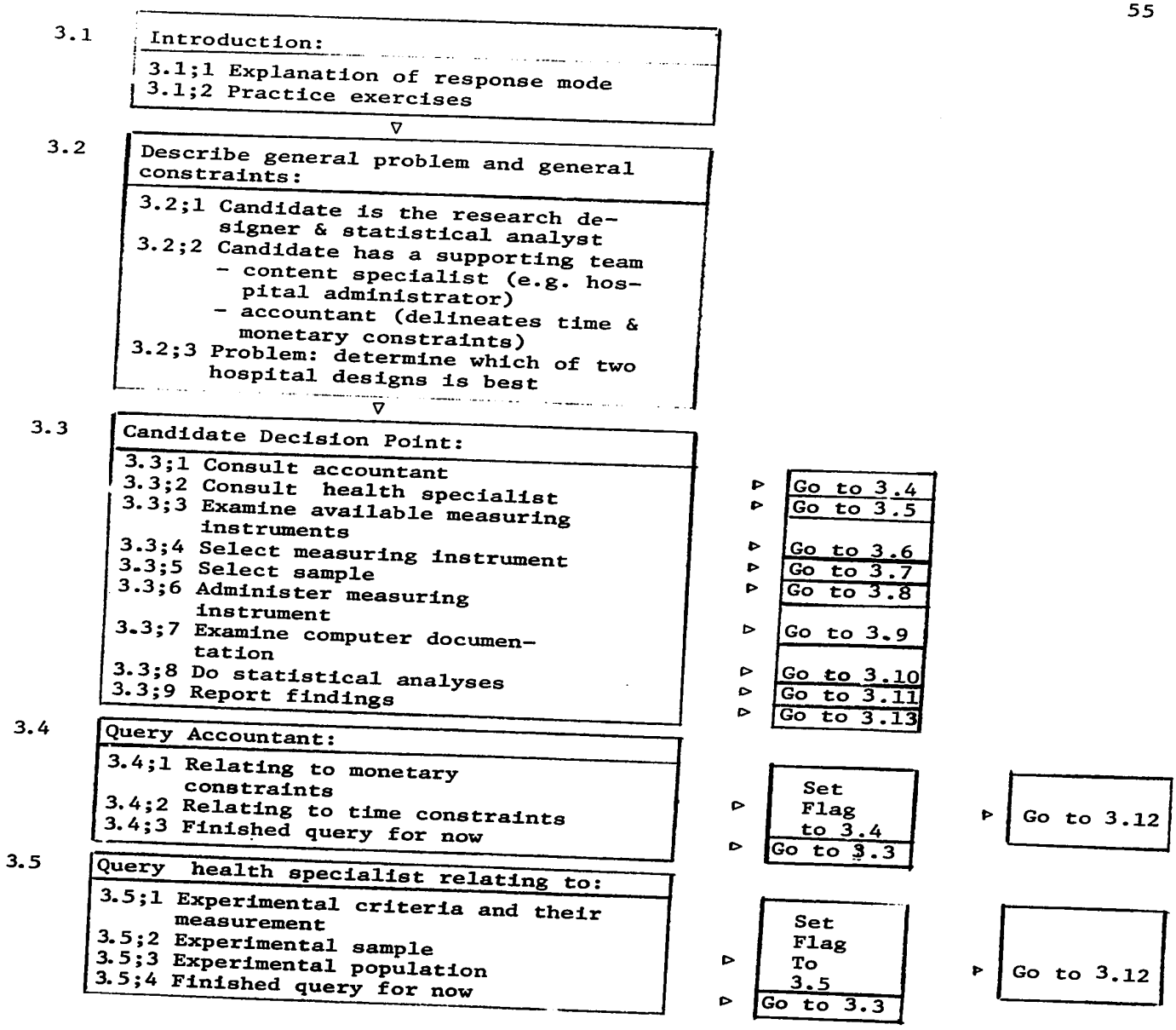
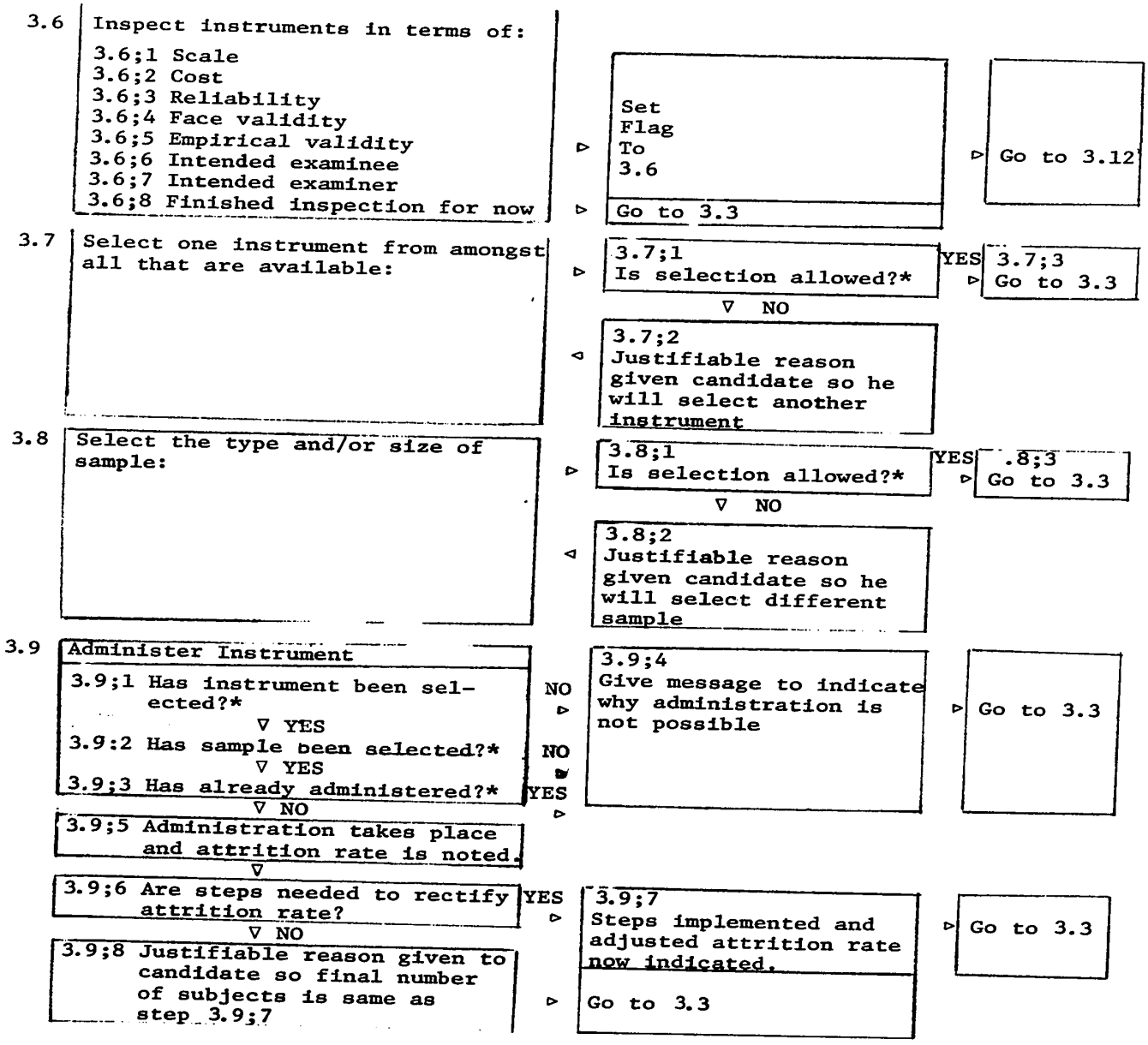


Figure 3

(3.1-3.14) Simulated Research Pathways



*Computer decision point

Figure 3 (continued)

(3.1-3.14) Simulated Research Pathways

After composing the essays and their keys, the author independently submitted the documents to two specialists, (both having similar qualifications to the specialists used in classifying multiple choice questions). After a series of iterative modifications the final documents are those given in Appendix B.

The reader will note that both essays have delineated the topics around which a subject's answer was to be composed. This was done since research (cf. p. 23) had indicated that all subjects should "run the same race"; that is to say, all respondents should have the same idea as to what was required in an answer. Furthermore, both keys were assumed to reflect the delineated topics in terms of factual, comprehension, and problem solving behavior. Factual marks were obtained if the subject simply used the terms specified in the two factual keys. No penalty was incorporated into the factual score if these terms were used incorrectly, though marks were lost in the comprehension section if incorrect use was evident. Accordingly, the key and the scoring scheme were assumed to reflect the behavioral characteristics previously defined (cf. p. 6-14). The problem solving score reflected the most complete answer, characterized by sophistication and judgement in the area of applied statistics and research design. For example, essay one describes a research problem which requires no inferential analysis since the entire population has been tested. Accordingly, demonstration of problem solving behavior was assumed to exist if the subject realized this and accepted the second hypothetical researcher's analysis as the correct procedure. The key for essay one describes further accepted behavior assumed to be characteristic of problem solving skill, and similarly for essay two and its key.

A fifteen minute introduction (Figure 3, step 3.1) was provided for subjects in order to (1) reduce the novelty effect of taking a simulation examination, (2) give subjects some practice in using the IBM 1500 terminals, and (3) define for the subjects the meaning of the terms "significant", "relevant", "irrelevant". These terms were used at specified points in the simulation (step 3.12;1) as choices for describing information they collected. "Significant" information described data that were absolutely necessary for properly analyzing the research problem. "Relevant" information described data that was informative but not absolutely crucial to an adequate solution, and "irrelevant" described unrelated, unnecessary information. The use of these terms will become more evident as the reader is directed through the steps of Figure 3.

After the introduction, general instructions (step 3.2) were given to the candidate. A situation was described in which he was to regard himself as a behavioral research consultant and statistical analyst, commissioned to answer this question: "Which hospital design is best--type one (traditional) or type two (spoke)." The candidate was told that a health expert and an accountant were being assigned to his staff as aids. The former, because of his specialized knowledge in the area of health, could be used as a source of information relating to valid health criteria. The accountant had the role of making sure the candidate stayed within the monetary and time constraints set by those who commissioned the study.

The following explanation will illustrate why a hypothetical health expert was used in these simulations. Most subjects had little or no acquaintance with the health care professions. Consequently, the situation

was novel ((cf. pp. 11-13) for definition of problem solving trait) and final solution of the problem would demand behavior which was characteristic of problem solving skills. Subjects could, however, interrogate the health expert (steps 3.3;2 and 3.5) in terminology that was very familiar to them (cf. pp. 6-8 for definition of factual trait). For example, some of the questions subjects could ask the health expert were:

(1) (a) What is a spoke designed hospital?
(b) What variable determines best design?
(c) What is the criterion of best design?

, and/or

(2) (a) What is a reliable measure of the criterion?
(b) What is a consistent measure of the criterion?

,and/or

(3) (a) What is the population to which inferences are to be made?
(b) What is the socio-economic status of the clientele in the population?

This author assumed that the subjects were very familiar with terms such as "variable", "criterion", "reliable measure", "social economic status", and so forth. As the subjects worked through the simulation, therefore, the need (or lack of need) for asking these questions was regarded to be factual behavior. Furthermore, the audio answers played back to these questions (step 3.12;1) were also worded in a form that was more research oriented than health oriented. For example, the audio answer to above question, numbered 1 (b), was

> In the minds of health officials, the variable that will determine which hospital design is best, is the degree or level of care patients receive. For example, if spoke design hospitals provide a higher mean level of care than traditional hospitals, then a spoke design would be considered the best architectual design to use for any hospitals constructed in the future.

Consequently, if the subject asked the standard questions in any behavioral research--what is the experimental criteria, can a random sample be drawn, to which population are inferences to be made, etc.--the answers he received, if understood, indicated which additional questions were or were not needed.

To determine if the subject did understand his answers, he was asked to classify each piece of information as being either significant, relevant, or irrelevant (step 3.12;2). If (and only if) he thought it very pertinent (i.e., significant) he was also asked an additional question (step 3.12;3) to determine if indeed he understood the significance of that information. For example, the answer to the question regarding "To which population are inferences to be made?" was: "all spoke and traditional design hospitals in Alberta." The subject who recognized this information as significant, was also asked the following question.

> The following categories of Alberta hospitals contain only spoke and traditional designs. To which hospitals will inferences be applicable once this study is completed?
> Rural hospitals
> Urban hospitals
> Hospitals with more than 75 beds
> Hospitals with less than 75 beds
> Hospitals with specialized doctors
> Hospitals with non-specialized doctors

Since the answer had indicated that inferences were to be made to all Alberta hospitals with the two architectual designs, the candidate who understood the answer would have indicated that inferences were applicable to all of the above. Such understanding was classified as comprehension behavior.

The candidate's querying of the accountant (steps 3.3;1 and 3.4), and his inspection of the available measuring instrument (steps 3.3;3 and

3.6) took the same format as that just described for consultations with the health expert (except that visual (not audio) answers were displayed (step 3.12;1) for all information relating to the measuring instruments). Likewise, the candidate was also asked to classify collected data according to pertinency (step 3.12;2) and if, and only if, he regarded the data as significant was he asked an additional question (step 3.12;3) in order to determine if he indeed understood why the data were crucial.

The hypothetical accountant served a useful role by controlling candidate performance without destroying the simulated real life experience. For example, when the candidate chose his measuring instrument (steps 3.3;4 and 3.7) or drew his sample (steps 3.3;5 and 3.8), mistakes could have been made if he had inadequately collected or had misunderstood previous data. The same was true if the subject did not take steps to reduce attrition rates in his sample (steps 3.9;5 - 3.9;8) or chose to do an inappropriate statistical analyses of his collected data (steps 3.11;5 - 3.11;6). In the last case, for example, the accountant gave a message that extra statistical analyses were done (in addition to those the candidate had requested) because the candidate's research had been monetarily efficient enough to allow for extra money to spend on additional analyses. This guiding function of the accountant accomplished two necessary steps for this study. Firstly, it reduced the potential immensity of such a program by allowing all candidates to access the same data which affected branching points. If no limitations are used then each consecutive branching point must have an increasing number of optional paths with convergence to a common point being realized only at the end. In Figure 3 the candidate was allowed to do anything within the specified pathways. If,

and only if, that attempt was deemed a posteriori to be particularly deviant, controls were enforced. Furthermore, because the accountant specified a "justifiable" reason for each control, it was hoped the attempt to approximate "real life" via simulation was not seriously reduced. To illustrate these above points an example is given from one of the simulations. The candidate was allowed to chose various hospitals as his sample (step 3.8). Some constraints were imposed (step 3.8;1) in that no more than one hospital of each design was allowed to be selected. The reason given was "Due to time and financial constrainst this study can not afford to test patients in more than two hospitals." Within this limitation, the subjects had an unrestricted choice among available hospitals, each of which had different numbers of beds. In order that statistical results did not have to be programmed for each possible pair of hospitals, attrition rates arose after testing (step 3.9;6 - 3.9;8), a common phenomenon in behavioral research. Attrition rates were such that (regardless of the size of hospitals any candidate had previously selected) all had the same number of completed measuring instruments. Consequently, the use of properly construed controls, can minimize the variety of paths that are possible without adversely affecting plausibility of the simulated experience.

The second reason for using the accountant as a controller was to enable a common comparison among subjects. For example, many of the statements to be or not to be selected in the report (step 3.13;2) were based on certain statistical analyses. If convergence in paths was not incorporated at step 3.11;6 it would not have been possible to ask all candidates the same questions in the report. To the degree that higher trait levels are dependent upon lower cognitive skills, assigned scores in the final report (problem solving) should have been dependent upon previous pathways (factual and comprehension). This assumes, however, that previous pathways were sufficiently representative of all factual and all comprehension behavior, an assumption which this author did not want to make. He did make the assumption, however, that controlled pathways previous to step 3.13 would not be a sufficient condition for the correct execution of step 3.13;2.

Figure 4 is a succinct illustration of the major controls used in the simulated problems. Numbers in this Figure correspond to steps previously described in Figure 3. Notice that all candidates were allowed to move at will between Point A and Point B. If execution in certain steps of Point B was not allowed, the candidate was taken to Point C where respective justifiable reasons were given so that the candidate would change his behavior in Point B before returning to Point A. Consequently movement to Point D

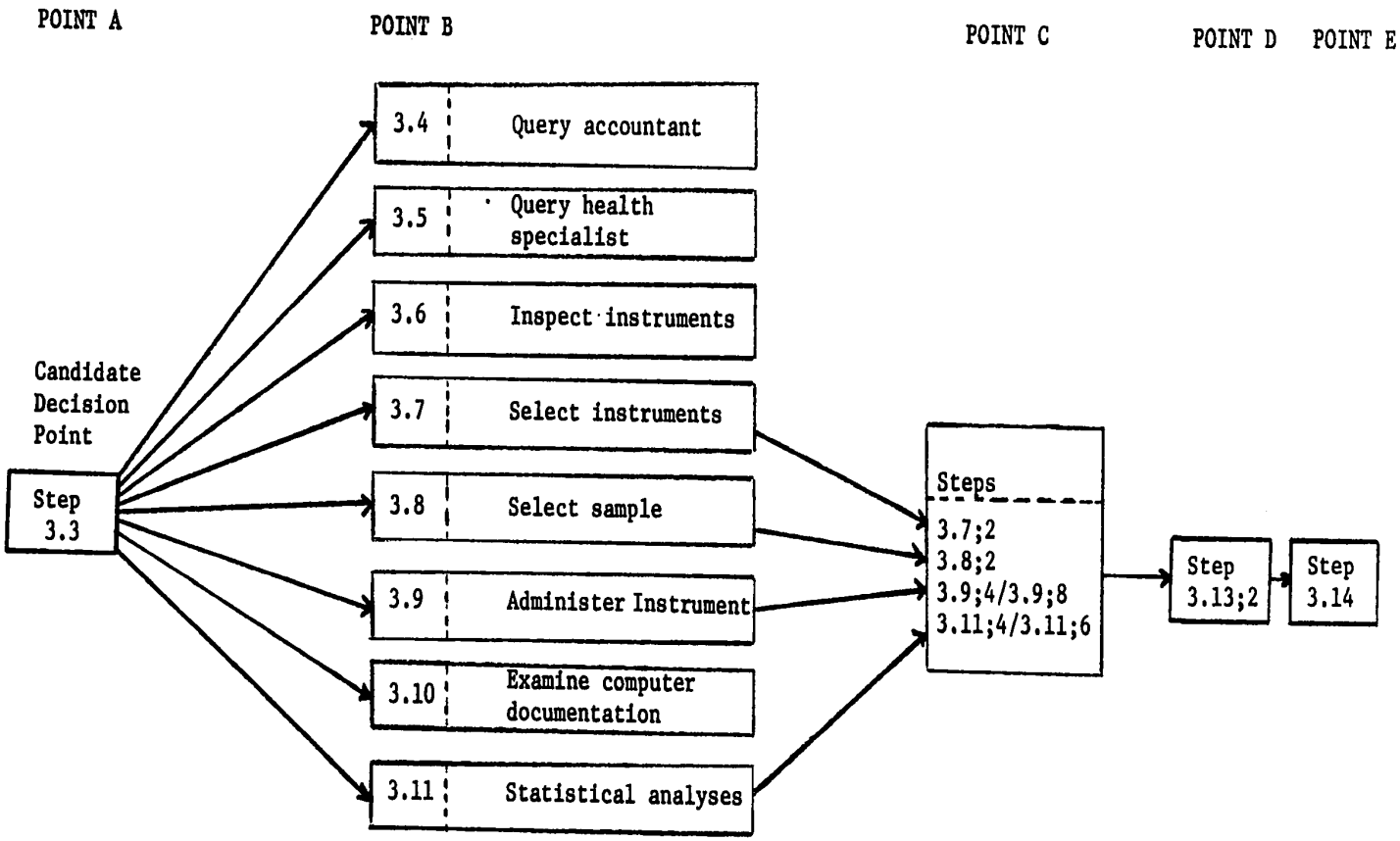


Figure 4: Controlling Pathway Variety in Simulated Problems
(Identified steps correspond to Figure 3)

took place only after all candidates had encountered similar experiences.

This particular subsection has dealt with a description of the paths available in the simulation problems used in this study. It should be noted that the simulation programs were pretested on only six subjects. The above explanation includes the modifications that were made as a result of information gathered in this pretesting process. The one variable that was not modifiable, but very present in the pretesting, was that these particular subjects viewed the simulation tasks as a measure of this author's ability in the area of statistical analysis and research design. Consequently this phenomenon may well have been present in the actual testing sessions, and as noted before (cf. p. 18),may be a source of invalidity in this study. The above explanation was given to acquaint the reader with a rather novel evaluation device. Due to this novelty, standardization among simulation programs has not yet been finalized. Consequently the analysis and interpretation of simulation data (cf. Chp. 4) may not be applicable to simulation programs in general. The reader is cautioned to view the results given in Chapter Four in terms of the above simulation characteristics. Before turning the reader's attention to those analyses, however, the following subsection describes the manner in which performances were scored and analysed for each trait level in the simulation programs.

Trait Classification and Scoring

It is to be noted that while 17 (nearly 20%) of the multiple choice questions had not been easily classified into one of the three trait categories, almost 100% unanimity was present in the trait classification process for the simulation exercises. This agreement may have been partially due to the fact that each of the two specialists was taken through the programs with this author present. Since the intent of each entry was explained at the time the specialists were making their classifications, results may have been biased in favor of unanimity. In any case, the only differences that were originally present were in the area of score weighting.

The factual questions in Figure 3, steps 3.4-3.6 fell into four categories: (1) those that were necessary, (2) those that were necessary but because of questions in category (1) were now redundant, (3) those that were informative but not fundamentally crucial to proper execution, and (4) those that were not needed and/or informative. Some examples will more precisely describe these four categories. On page 59 of this chapter some questions a candidate could ask the health expert were quoted. Statements 2(a) and 2(b) actually mean the same thing and consequently equivalent replies were given to both questions by the computer. Since a behavioral researcher would want to know what was a reliable (i.e., a consistent) measure either one of those questions should have been asked. If asked, marks contributing to the factual component of the simulation score would have been assigned. If, however, both questions had been selected it is reasonable to assume the candidate did not know that "reliable" and "consistent" mean the same thing (that is, the candidate did not display an ability of "knowledge of terminology" (Bloom, et al,

1956, pp. 64-65)). By selecting both, however, it was assumed that the candidat saw the need for collecting data on "reliable measures" or "consistent measures". In order to discriminate among candidates no marks were assigned if neither of the questions was asked; one mark was assigned if both questions were asked; two marks were assigned if only one question was asked. In essence the scoring scheme assigned one point to the factual score for recognizing the need for such information and one point for remembering the definitions of "consistent" and "reliable". For another example of necessary and redundant information see statements 1(b) and 1(c) (cf. p.59).

The reader's attention is now directed to an explanation of questions that were informative but not fundamentally crucial to proper solution of the problem. On page 59, statement 1(a) was so classified. A definition of "spoke designed hospitals" was not needed for solving this problem; it was simply a hypothetical experimental unit. However, subjects were not well acquainted with the health area and they may have perceived this question to be necessary. All questions that were informative but not fundamentally crucial to final solution were assigned a weight of zero.

The last type of question--that which was unnecessary--is illustrated on page 59, statement 3(a). If a candidate had asked the most important question, namely 3(a), he would have found out that 3(b) was irrelevant. That is because the population was actually a population of all hospitals in Alberta, statement 3(b) had become meaningless. Bloom, et al (1956) speak of "knowledge of the techniques and methods used by scientists in seeking to answer questions about the world [p. 74]". In research one of the first and most basic questions the designer must ask is that relating to the

inferential population. It was assumed, therefore, that factual behavior would be evident if (1) statement 3(a) was asked, and (2) that this question was the first question to be asked. A score to reflect these latter two categories was accomplished in the following manner. Zero was assigned if neither 3(a) or 3(b) was selected, or if only 3(b)--the unnecessary statement--was selected. One mark was assigned if both 3(a) and 3(b) were selected. Two marks were assigned if only 3(a)--the crucial question--was selected. It is to be noted that while penalties were used for inefficient selection of data, a candidate could never obtain a negative score. This was assumed to be reasonable because asking too many questions can rarely hurt a designer, only make him less efficient. Furthermore, few (if any) of the questions were blatantly obvious as being unnecessary on first inspection. They were assumed to be so only if the subject had asked the most fundamental questions inherent to research design (i.e., he had exhibited knowledge of methodology). The use of scores less than zero, however, would mean that a lack of factual behavior in knowing methodology takes away from the degree of factual behavior in the area of definitions (for example). Consequently, the scoring scheme reflected that a candidate either did (positive score) or did not (zero score) demonstrate mental behavior described as factual.

In the above explanation the reader will note that a simple algorithm can describe the scoring scheme (cf. Figure 5). This algorithm was used for scoring each set of questions in steps 3.4-3.6 and the results of each set were added together to obtain the candidate's factual score. However, the final factual score also included marks for knowledge of methodology outside steps 3.4 to 3.6 in Figure 3.

1. Candidate's obtained score: $Wa - b - d^*$

   (i) if $a=A \wedge b=0 \wedge d=0$ then maximum score obtained

   (ii) if result is negative set score to zero for any one set of questions in steps 3.4 or 3.5 or 3.6 of Figure 3.

(N.B. this scoring scheme was also used for assigning comprehension scores in steps 3.10 and 3.11;5 (cf. p. 71)).

*where

N is total number of questions in any one step of Figure 3

A is total number of "unique" necessary questions

a is total number of selected A's

B is total number of "redundant" necessary questions

b is total number of selected B's

C is total number of informative but not necessary questions

D is total number of unnecessary questions

d is total number of selected D's

W is weight of any one unique necessary question defined as: $\frac{N-B-C}{A}$.

2. Example:

In a total of 10 questions (N=10), two are uniquely necessary (A=2), two are redundantly necessary (B=2), two are only informative (C=2), and four are unnecessary (D=4), then the weight for each unique necessary question is 3: $W= \frac{10-2-2}{2}$.

Subject selecting both unique necessary questions (a =2), one redundant necessary question (b=1), and four unneccessary questions is given a score of 1: (3(2)-1-4).

Figure 5: Scoring Algorithm

Cognitive behavior was classified as factual if the candidate chose to administer the measuring instrument (Figure 3, step 3.9) after he had selected his instrument (steps 3.7), after he had drawn his sample (step 3.8), and provided he had not previously administered the measuring device. (These particular checks for the order of execution are specified in Figure 3, steps 3.9;1 - 3.9;3). If the candidate did not receive any messages indicating his execution order was impossible (step. 3.9;4) he was rewarded factual marks. Similar marks were awarded at steps 3.11;4 and 3.13;3.

The classification and the assignment of weighted scores for comprehension will now be explained. As previously described (cf. p. 59) candidates were asked if each piece of information they collected in steps 3.4-3.6 was either significant, relevant, or irrelevant (step 3.12;2). Classifying information into one of these categories was considered to be characteristic of comprehension behavior. Bloom, et al (1956) might describe this as "The ability to interpret various types of...data [p. 94]" (i.e., interpretation); but he might also describe it as the "ability to recognize what particulars are relevant...[p. 147]", (i.e., analysis). This equivocation in Bloom's classification is reduced by the use of this study's trait called comprehension (cf. pp. 8-11).

Since a perfectly efficient candidate would not select any information that was not significant, comprehension scores were assigned only in step 3.12;2 following the selection of unique and necessary questions (such questions are designated as "a" in Figure 5). If the candidate designated such information as irrelevant no marks were assigned; marks of one and two were assigned to selections of relevant and significant respectively.

For correctly selecting the category of significant, an additional question was asked (step 3.12;3); an illustration of such a question has been previously given to the reader (*cf. p. 59*). If the candidate correctly answered this question an additional mark was assigned to his over-all comprehension score. These questions attempted to elicit behavior in using or "applying" (Bloom, et al, 1956, p.124) the information a subject had just collected. Comprehension scores were also assigned in steps 3.10, and 3.11;5. In the former step the candidate inspected documentation of various statistical programs that might be used for analyzing his data, and in step 3.11;5 he actually requested such analyses. Possible selections in these two steps were classified in an identical manner to those in steps 3.4-3.6, namely, some were necessary, others were necessary but redundant, etc.. Consequently the scoring scheme of Figure 5 was used to calculate these particular comprehension scores. Finally, marks assigned in steps 3.7 and 3.8 were also classified as comprehension. It was assumed the proper selection of an instrument (step 3.7), or a sample (step 3.8), required more than factual behavior, because these selections were dependent upon much of the information collected in steps 3.4-3.6. However, if the accountant had to control any of these selections (steps 3.7;1, 3.8;1, and 3.12;6) no marks were assigned because the first selection had been in error. The total comprehension score for any candidate was a simple sum of all assigned scores in the steps previously designated as comprehension.

Before dealing with the scoring scheme of problem solving behavior, the reader should note that the candidate was allowed to move at will between

steps 3.3-3.6 and 3.10. The constant return to the major decision point (step 3.3) allowed the candidate to review or collect additional information. The scoring scheme described above for factual and comprehension behavior was applicable only to the first selection of any one question. Consequently review did not increase or decrease scores unless a new question was being asked in any one of the steps between 3.3-3.6 and 3.10.

To properly complete the report of his findings (step 3.13;2) it was assumed that the candidate had to exhibit problem solving behavior. A series of 29 statements (each assigned a weight of one for scoring purposes) was presented to the candidate. He was required to either select or reject each statement. Selection indicated that he considered the statement to be valid and correct (based on his findings and interpretations of collected data). Statements in the report dealt with experimental limitations and statistical interpretations. If the candidate wished to review his computer analyses he was provided with these before making his selection. Except for the effect of guessing, correct selection was considered to be dependent upon the candidate's understanding of all information he received in prior steps (3.4-3.12). For example, one simulation developed a situation in which the total experimental population (defined to be of finite size) was tested. Consequently mean differences were automatically "significant" in the statistical sense. Blind use of computer programs which calculated means and probabilities of differences of means based on populations of infinite size would seriously limit the correct selection of a statement such as this:

It is reasonable to state that the means in the EXPERIMENTAL POPULATION are DEFINITELY DIFFERENT.

Harasym (1970) has hypothesized that the measurement of higher cognitive abilities is difficult in multiple choice examinations because examinees can answer such items by simply reducing unlikely alternatives. This behavior probably is a lower cognitive skill than a skill which is used to answer the stem and then finds the corroboration within the answer set. Accepting this assumption, this author programmed statements for the report so that the candidate saw only one statement at a time. Furthermore, the candidate was allowed only one chance to include or not include any one statement in his report. The sum of all the correct selections in the report was used as a candidate's total score for problem solving.

The above scoring schemes for factual, comprehension, and problem solving behavior were programmed into the IBM 1500 system so that gross scores for each trait were automatically given for each subject. The scoring routine was thoroughly debugged and tested so that this author is confident that perfect accuracy was obtained. This author highly recommends the use of programmed scoring routines. For example, test runs on the scoring scheme often indicated that the score manually obtained was different from the computer score. After careful re-examination, however, the manually calculated scores were found to be incorrect and the computer scores exactly correct. Furthermore, the sheer volume of printed output for any one candidate's execution makes it too difficult for an examiner to properly analyze the results without an automated scoring routine.

The analysis reported on trait scores in the simulation tasks (cf. Chp. 4) has been restricted to estimations of consistency between alternate forms and validity estimates via Campbell and Fiske's multitrait multimethod matrix. Item analyses have not been reported on individual scores that were added into over-all trait scores for subjects, simply because their extraction from individual's pathways was too difficult. The average length of time spent by subjects in the two simulations was 124 minutes. Inspection of pathways taken by various candidates indicated that stereotype movement through the simulations was not present. Furthermore paths were often retraced in steps 3.4-3.6 for purposes of review or for purposes of collecting additional data. Tracing such paths in order to manually assign individual marks to individual selections is most difficult and often inaccurate. An item analysis contributes minimal information to a study of this kind, and consequently the need for such an analyses did not justify the effort required to extract data for it.

Up to this point, this chapter has described how the trait scores were classified, scored and analyzed for multiple choice, essay, and simulation tasks. In addition a description of the simulation exercises was provided for the reader in order that the use of this novel technique (in this particular study) was clearly understood. A listing of these simulation programs is too lengthy for inclusion in this report; instead the standardized documentation for programs on the IBM 1500 CAI system at the Division of Educational Research Services (University of Alberta) is supplied to the reader in Appendix C.

The previous sections in this chapter were devoted to the ways and means of empirically estimating construct validity by examining candidate behavior in various testing methods. As added information for this study subjects were also asked to supply information for the establishment of face validity of these same methods and traits. The following section is a description of that data.

Establishment of Face Validity

Levine, McGuire, and Nattress (1970) have used a questionnaire to determine how candidates perceived various forms of examinations. This investigation used three of the items from this questionnaire (see Appendix D, questions 1 to 3) to assess examinees' feelings toward all three forms of examinations (essay, multiple choice, and simulation) in terms of their difficulty and relevancy. This study also used part of the Class Activities Questionnaire (CAQ) developed by Steele (1969, 1970). Though the CAQ was originally designed for assessing the cognitive, behavioral, and affective intent and practice of classroom instruction, some of its items (with slight rewording) appeared to have relevance for assessing the face validity of examinations used in this study. Steele has shown that different pairs of items load well on the first seven factors which he interprets as corresponding to Bloom's taxonomy. Provided the rewording of the original statements did not seriously change these loadings it was assumed questions 4 and 9 load on a memory factor, questions 7 and 13, 6 and 12, 8 and 11, 10 and 15, and 5 and 14 load on factors of interpretation, application, analysis, synthesis and evaluation respectively.

Subjects filled in these questionnaires according to the instructions provided in Appendix D. In a similar manner to the descriptive analysis that Steele (1969) did with his questionnaire, this study used only a cross-classification of dichotomized pairs of items. Accordingly, the percentages of agreement between pairs of items only are reported. The particular standards set for establishing meaningfulness of these data were more demanding than those used by Steele (1969). Firstly, at least 67% of the subjects had to give the same answer to both items (in any respective pair) before any further inspection was made of the data. In other words, items were considered unreliable for measuring a particular trait-method variable if 67% consistency was not obtained. Secondly, 67% of the subjects had to agree that a particular method did indeed have the face validity for measuring a particular trait. For this analysis see Chapter Four, pp. 125-130.

Nature of Sample and Test Conditions

As previously stated (cf. p. 8) 50 subjects were tested in this study. Seventeen of these 50 were taking their first course in applied statistics and experimental design, 10 of the 50 were taking their second course in this same content area, and the remaining 23 subjects had completed at least these two courses. For future reference the group of 17, 10, and 23 subjects are labelled group 1, group 2, and group 3 respectively. The first 17 subjects (group 1) were tested after they had completed about five-sixths of their introductory course; the second group of 10 subjects were tested after they had completed about one-half of their

second course. The majority of the remaining 23 subjects (group 3) sat with these 10 subjects when writing the tests; the remaining proportion of these 23 subjects sat in the examination sessions given for group 1. Chapter Four will (cf. pp.123-125) provide an analysis of expected direction of mean differences for these groups, as further evidence in establishing construct validity for factual, comprehension, and problem solving traits.

The subjects were told the nature of the study and therefore were aware that the examinations were attempting to measure the traits previously described (cf. p. 6-14). Operating under the assumption that test anxiety would interfere with these behaviors, this author assured each subject that his performances would be known only to this author and himself. To enhance this anonymity all respondents were given a randomly selected identification number to be used by the subject. Coffee was also available in the written sessions for the multiple choice and essay tests in order to further reduce test anxiety. In order to help motivate subjects, however, the author encouraged subjects to view these tests as practice sessions for their courses and/or upcoming oral examinations. A set of tests (multiple choice, essay, and simulation) was administered in a 2-1/2 hour block between 3 P.M. and 5:30 P.M. on one day and the second set of examinations was administered at the same time the next day. The three groups described above were randomly divided into two sections; the first section wrote the multiple choice and essay examinations while the second section took the simulation test in another room. When individual subjects finished their respective tests

they proceeded to the next room and finished their testing for that day. The identical procedure was repeated the following day.

Since reliability estimates were required to reflect consistencies over time and alternate method, no alternate forms were allowed to be administered in one day. The two multiple choice tests, as well as the two essays and two simulation programs had been arbitarily assigned numbers 1 and 2. Due to the fact that one simulation and one essay dealt with descriptive statistical analyses, a decision was made to administer these tests on the same day. Since groups were randomly assigned into two sections (each of which would take either the essay or simulation first followed by the second test) it was assumed that advantage of a particular method-order would be confounded in any analyses of all test behavior. As it turned out it was essay one (E_1) and simulation two (S_2) that dealt with descriptive analyses. A random selection from a hat was used to make the decision that multiple choice one (M_1) would be administered in the same block as the above two tests. Thus, multiple choice two (M_2), essay two (E_2), and simulation one (S_1) constituted the second test session.

Another random selection from a hat decided the order of tests within any testing session that each section would take. As it turned out, the first section took the multiple choice, then essay, then simulation; the second section took the simulation, multiple choice, and essay. In order to facilitate administration and inter-room movement for both subjects and the administrative help used by the author, the above procedure was repeated the second day.

Finally, because four different days were needed to complete testing on all 50 subjects (a restriction imposed by the number of terminals in the IBM 1500 system) a randomization of blocks of tests was also done. Selecting from a hat once again, it turned out that the subjects taking the introductory course (group one) plus those in group three who were taking the examinations with group 1 took the block of tests labelled (M_2, S_1, E_2) on the first day and the second block (M_1, S_2, E_1) on the second day. Those in group 2 and 3 who sat together took these blocks of tests in reverse order.

This section has described the nature of this study's sample, test conditions, and procedures for determining testing order. The reader will be aware that the randomization of testing orders within blocks of examinations and across groups of subjects was of limited value because the number of possible selections was extremely limited. The above procedures were done, however, in the hope that order of testing as well as "testing effects" (Campbell and Stanley, 1963, p. 5) might be confounded in any analyses of the data.

Summary

This chapter has dealt with the experimental design of this study. The reader was first introduced to Campbell and Fiske's (1959) model for establishing reliability and validity estimates. Since Campbell and Fiske's model has certain limitations, Joreskog's (1967) factor analytic procedure for a restricted maximum likelihood model was described as a reasonable means for succinctly analyzing the multimethod-multitrait matrix.

The particular manner in which trait scores were developed, scored, and analyzed in each of the methods has also been described. An attempt was made to show how this author and two additional specialists classified questions and behaviors required to answer them in terms of the described traits called factual, comprehension, and problem solving.

The KR_{20} formula was described as a reasonable means for determining inter-item consistencies in various subtests of the multiple choice test. The Anova model was put forward as the most plausible means for establishing interjudge consistency in essay marking. The modified CAQ questionnaire was selected for estimating the face validity of all tests.

The final topic covered in this study's experimental design section was that describing the nature of the sample and testing conditions. Approximately half of the 50 tested subjects were (at the time of this study's collection of data) taking courses in the content area of applied statistics; the other half had completed at least two of these courses. Testing was done under non-stress conditions, except that subjects were expected to finish one set of three examinations within a time period of two and one-half hours. All three methods and their alternate forms were administered over a two day period, but the order in which they took these examinations was assumed to be random. Complete randomization was not possible, however, because this author had imposed the constraint that no alternate forms could be administered in the same day.

CHAPTER FOUR

DATA ANALYSIS AND INTERPRETATIONS

Introduction

This chapter contains the results of the application of the various statistical analyses described in Chapter Three. The reporting of these results follows the same order as described in that chapter: (1) Campbell and Fiske's multimethod-multitrait matrix for estimating convergent and discriminant validity as well as alternate form reliabilities, (2) Jöreskog's restricted maximum likelihood estimates based on the simultaneous use of all correlations in the above matrix, (3) item analysis of the multiple choice items, including inter-item consistency estimates, (4) inter-judge consistency estimates for essay markers, (5) expected mean differences amongst subjects on each subtest, and (6) face validity estimates of each method for each trait. Each of the above six topics will be dealt with in the following manner: (1) a report of the findings, and (2) a discussion and/or interpretations of those findings. Chapter Five will deal with the general implications of these findings as they relate to educational measurement and decision-making.

Analysis of the Multitrait-Multimethod Matrix

Introduction

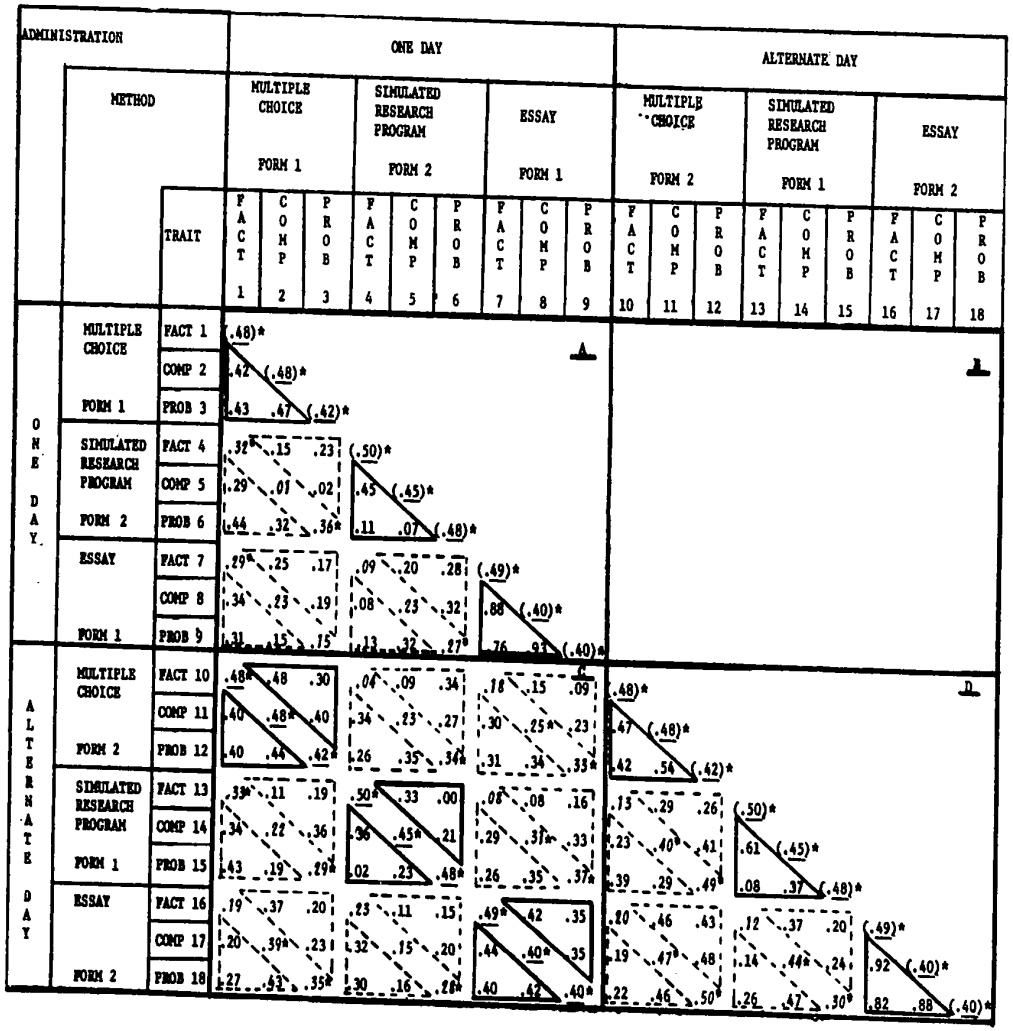
As previously described and illustrated (cf. pp. 33-35), the administration of two alternate forms of each method enabled estimates to be made of reliabilities reflecting consistencies of alternate forms over a time period of one day, as well as estimates of convergent and discriminate validity. Contained in Table 1 are the intercorrelations resulting from the administration of two blocks of tests--M_1, S_2, E_1, and M_2, S_1, E_2 on two different occasions. This table conforms to the model illustrated in Figure 1, page 34 . The submatrices in Figure 1 (A, B, C, D) have been similarly designated in Table 1. A contains the intercorrelations of one battery of tests administered on one day; D contains similar estimates for the second battery which also was administered on one day. Submatrix C and B contain identical intercorrelations, reflecting consistencies of time, form and/or traits.

The reliability estimates in Table 1 ($r_{10,1}$=.48; $r_{11,2}$=.48; $r_{12,3}$=.42;...; $r_{18,9}$=.40) have been underlined. These values have also been used to replace the "1's" in the main diagonals of A and D. For readability the convergent validity values in A, C, and D have been typed in italic script.

The reliability and validity estimates in Table 1 were subjected to statistical tests of significance, to estimate respective

TABLE 1

MULTITRAIT-MULTIMETHOD MATRIX



*Significant at p=.05 (one-tailed t test)

probabilities of being different from zero. Since theoretically reliability is defined as a ratio of variances, and since practically speaking convergent validity should be positively correlated, a one-tail "t" test was the appropriate means for estimating the likelihood of population values (ρ) being zero. This author chose to set his level of significance (α) at 0.05. The critical value of t (with this study's 48 degrees of freedom) for rejecting a null hypotheses of $\rho=0$ was 1.673. Accordingly, $_{.05}t_{48}$ required a correlation of at least 0.235 before the assumption was made that the observed estimate was drawn from a population characterized by $\rho\neq0$. Under this criterion the reliability and convergent validity values which are regarded as significant have been designated with a star (*). Note, however, that many other intercorrelations in Table 1 which this study regards as neither reliability nor convergent validity values would also be considered as significant. It should be remembered that while tests of significance are a useful step in making reliability and validity estimates in this study, they do not provide sufficient information for estimating the worth of those estimates which are probably not zero.

Reliability

Reliability estimates in this study were the estimated consistencies between alternate forms of the same method measuring each trait. Values range from $r_{13,4}=.50$ (correlation between

factual subtests in the two simulation programs) to $r_{18,9}$=.40 (correlation between problem solving subtests in the two essay questions). Inspection shows that the nine reliability estimates are (1) very low, even though significantly different zero, and (2) all are quite comparable in terms of degree of consistency. Efforts aimed at keeping mental behavioral reactions to measured content the same in alternate forms has not been accomplished to any respectable degree. Theoretically, it is impossible for any monotrait-monomethod value (i.e., a reliability estimate established over a period of one day) to be less than any other correlation lying in its row or column of submatrix C because a variable's true score cannot correlate with a second variable's true score more than it does with itself. Inspection of Table 1: C indicates violations of this requirement are present.

Reliability Estimate	<	Correlation Value
$r_{12,3} = .42$	<	$r_{12,2} = .44$
$r_{17,8} = .40$	<	$r_{16,8} = .42$; $r_{17,7} = .44$
$r_{18,9} = .40$	<	$r_{18,2} = .43$; $r_{18,8} = .42$

Why did these theoretical violations occur? Firstly, reliability estimates as defined in this study are not correlations of a variable with itself, but rather correlations of alternate forms presumably measuring the same trait over a time period of one day. Consequently any of one of these reliability estimates "...is a little more like a validity coefficient than is an immediate test-retest reliability for the items are not quite identical [Campbell and

Fiske, 1959, p. 83]." Secondly, four of the above five correlational values which exceed their respective reliability estimates ($r_{12,2}$; $r_{16,8}$; $r_{17,7}$; and $r_{18,8}$) have common methods. Consequently, if trait influence was small relative to method influence, the possibility was enhanced for finding inappropriate higher correlations in any of the monomethod triangles in C.

On the other hand, if trait influence was greater than the influence of time and method, each reliability estimate in A, $r_{1,1}$=.48 to $r_{9,9}$=.40 (and identically in D, $r_{10,10}$=.48 to $r_{18,18}$=.40) should exceed all correlation values lying its row or column of its respective submatrix. The following conditions indicate that time and/or method may have been more influential than trait.

Reliability Estimate	<	Correlation Value
$r_{3,3} = .42$	<	$r_{3,2} = .47$
$r_{7,7} = .49$	<	$r_{8,7} = .88$; $r_{9,7} = .76$
$r_{8,8} = .40$	<	$r_{8,7} = .88$; $r_{9,8} = .93$
$r_{9,9} = .40$	<	$r_{9,7} = .76$; $r_{9,8} = .93$
$r_{11,11} = .48$	<	$r_{12,11} = .54$
$r_{12,12} = .42$	<	$r_{12,11} = .54$; $r_{15,12} = .49$
	<	$r_{16,12} = .43$; $r_{17,12} = .48$
	<	$r_{18,12} = .50$
$r_{13,13} = .50$	<	$r_{14,13} = .61$
$r_{14,14} = .45$	<	$r_{14,13} = .61$; $r_{18,14} = .47$
$r_{15,15} = .48$	<	$r_{15,12} = .49$
$r_{16,16} = .49$	<	$r_{17,16} = .92$; $r_{18,16} = .82$

$$r_{17,17} = .40 < r_{17,11} = .47;\ r_{17,12} = .48 < r_{17,14} = .44;\ r_{17,16} = .92$$

$$r_{18,18} = .40 < r_{18,11} = .46;\ r_{18,12} = .50 < r_{18,14} = .47;\ r_{18,16} = .82 < r_{18,17} = .88$$

It is to be noted that Campbell and Fiske do not suggest the immediately above comparisons. Presumably these authors feel such criteria involving time are highly rigorous and often violated. While later discussion on convergent and divergent validity will adhere to Campbell and Fiske's criteria, it should be noted that if test behavior was best described by the three traits, the influence of one day between test batteries should not be as influential as observed in the above comparisons.

Convergent Validity

Convergent validity estimates are contained in all offdiagonals in Table 1. The size of the offdiagonal elements in submatrices A and D have confounding influences of trait and method, those in C have confounding influences of trait, method, and time. Unlike the reliability tests of significance, all convergent tests do not lead one to assume that different methods attempting to measure the same traits are characterized by dependence. Indeed, of the 36 convergent validity estimates established in Table 1, 16 of these values may have corresponding population values of zero:

Table 1

	Subtests Correlated	
Submatrix A	Trait	Methods
$r_{5,2} = .01$	Comprehension	: S_2 & M_1
$r_{7,4} = .09$	Factual	: E_1 & S_2
$r_{8,2} = .23$	Comprehension	: E_1 & M_1
$r_{8,5} = .23$	Comprehension	: E_1 & S_2
$r_{9,3} = .15$	Problem	: E_1 & M_1
Submatrix D		
$r_{13,10} = .13$	Factual	: S_1 & M_2
$r_{16,10} = .20$	Factual	: E_2 & M_2
$r_{16,13} = .12$	Factual	: E_2 & S_1
Submatrix C		
$r_{10,4} = .04$	Factual	: M_2 & S_2
$r_{10,7} = .18$	Factual	: M_2 & E_1
$r_{11,5} = .23$	Comprehension	: M_2 & S_2
$r_{13,7} = .08$	Factual	: S_1 & E_1
$r_{14,2} = .22$	Comprehension	: S_1 & M_1
$r_{16,1} = .19$	Factual	: E_2 & M_1
$r_{16,4} = .23$	Factual	: E_2 & S_2
$r_{17,5} = .15$	Comprehension	: E_2 & S_2

It is to be noted that 15 of these 16 nonsignificant values are estimates established on either the factual or comprehension traits;

(9 were those established on factual, the remaining 6 comprehension). Or put another way, of the 12 estimates established for factual convergent validity, only 3 estimates were classified as probably significant; (see Table 1: $r_{4,1}$=.32 and $r_{7,1}$=.29, and $r_{13,1}$=.33). Similarly, of the 12 estimates established for comprehension convergent validity, only 6 estimates were classified as probably significant:

Table 1	Subtests Correlated	
Submatrix D	Trait	Methods
$r_{14,11} = .40$	Comprehension	: S_1 & M_2
$r_{17,11} = .47$	Comprehension	: E_2 & M_2
$r_{17,14} = .44$	Comprehension	: E_2 & S_1
Submatrix C		
$r_{11,8} = .25$	Comprehension	: M_2 & E_1
$r_{14,8} = .31$	Comprehension	: S_1 & E_1
$r_{17,2} = .39$	Comprehension	: E_2 & M_1

In the case of all three significant factual trait estimates, values are hardly "...sufficiently large to encourage further examination of validity [Campbell and Fiske, 1959, p. 82]", provided one assumes that at least 16% (r^2, where r=.40) of one subtest's variance should be predictable from the scores of the other subtest with which the first was correlated. This criterion while minimal in an absolute sense, is not unreasonable since at best only 25% of the

variance is predictable in those subtests that were correlated and then regarded as reliability estimates. Using this same criterion only the comprehension convergent validity estimates in submatrix D warrant any further inspection.

As previously indicated only 1 of the 12 convergent validity estimates for problem solving was not assumed to be significantly different from zero (see Tables 1; $r_{9,3}$=.15). However, using the above criterion for further inspection only those values in submatrix D designated as $r_{15,12}$=.49 and $r_{18,12}$=.50 justify the reader's attention. This is not to say that some variance is not accounted for by problem solving behavior in other significant correlations. It simply implies the degree of influence this behavior may have in this study's test measurements is not sufficient to justify the evaluators attention any further. Further inspection is that relating to discriminate validity which will now be discussed. By a process of elimination only 5 out of 36 estimates of trait validity need now be discussed in terms of this criterion.

Discriminate Validity

The need for discriminant validity is due to the fact that tests "...can be invalidated by too high correlations with other tests from which they were intended to differ [Campbell and Fiske, 1959, p. 81]." Therefore the five correlations described above as deserving further inspection should be higher than corresponding heterotrait-monomethod values as well as corresponding heterotrait-heteromethod values.

The reader's attention is first directed to those 3 comprehension validity estimates in submatrix D that warrant the establishment of discriminant validity, namely $r_{14,11}$; $r_{17,11}$; and $r_{17,14}$. As seen below not one of these convergent values exceeds all corresponding values in its row or column of the dotted heterotrait-heteromethod triangles in D.

Comprehension Values	<	Corresponding Heterotrait-heteromethod Values
$r_{14,11} = .40$	<	$r_{14,12} = .41$
$r_{17,11} = .47$	<	$r_{17,12} = .48$
$r_{17,14} = .44$	<	$r_{18,14} = .47$

It is to be noted that these differences are small, and furthermore, are the only such inappropriate differences that exist for the three validity estimates under discussion. As Campbell and Fiske (1959, p. 84) point out this is not an unusual phenomena in most psychological and educational measurements. The degree of relationship between variables that have neither method nor trait in common should of course be less than variables with common traits. If the essay, multiple choice, and simulation methods were independent, and similarly if factual, comprehension, and problem solving behaviors were unrelated, values in the heterotrait-heteromethod triangles in all tables should be close to zero. The fact that they are not is evidence that method covariance and/or trait covariance is present; and to the degree respective heterotrait-heteromethod values exceed

or are equivalent for the three convergent validity estimates under discussion, indicates sources of invalidity.

The reader should also be cautious when comparing the relative size of $r_{14,11}$=.40; $r_{17,11}$=.47; and $r_{17,14}$=.44. If method variance is high all correlations in the heteromethod block are elevated, <u>including</u> the validity diagonal. This is well illustrated in <u>D</u>, in particular for correlations $r_{14,11}$; $r_{17,11}$ and $r_{17,14}$. For example, the possible method variance in Essay 2 (see the solid monomethod triangle in the lower right hand side of <u>D</u>) may account for the fact that $r_{17,11}$=.47 (correlation of essay and multiple choice comprehension) is higher than $r_{14,11}$=.40 (simulation and multiple choice measures of comprehension.) Since method variance can inflate validity estimates, therefore, previous criterion of not inspecting any correlations less than 0.40 is even more justified.

Why should convergent validity estimates exceed all related values in the monomethod-heterotrait triangles? The requirement is obvious since a variable should correlate higher with an independent effort to measure the same trait than with measures designed to assess different traits but which employ the same method. As seen below the three convergent validity values which warrant inspection--$r_{14,11}$; $r_{17,11}$; and $r_{17,14}$ of <u>D</u> usually do not exceed their corresponding values in monomethod triangles.

Comprehension Convergent Values	<	Corresponding Monomethod-heterotrait Values
$r_{14,11} = .40$	<	$r_{11,10} = .47$; $r_{12,11} = .54$; $r_{14,13} = .61$
$r_{17,11} = .47$	<	$r_{12,11} = .54$; $r_{17,16} = .92$; $r_{18,17} = .88$
$r_{17,14} = .44$	<	$r_{14,13} = .61$ $r_{17,16} = .92$; $r_{18,17} = .88$

Absolutely no validity estimates involving essay measures, exceed even one corresponding inter-trait correlation in any essay monomethod triangles in A or D. It is possible high intercorrelations among factual, comprehension, and problem solving traits accounts for high values in most monomethod triangles. However, monomethod triangles for both essays indicate very high intercorrelations among traits, while the two simulation methods show some separation of traits; the two multiple choice methods fall in between these two ranges. Since some separation among traits was possible in simulation (except for factual-comprehension intercorrelations) it is reasonable to infer method variance contributes to the observed high intercorrelations in both the multiple choice and essay methods. Furthermore, the high intercorrelations observed between factual and comprehension traits in the simulation programs seems (in retrospect) to be very reasonable, since the scores in the factual subtest of both simulations was dependent upon the candidate's understanding previous information (cf. pp. 59-60). That is, some information was classified as unnecessary if he had asked for and

understood previous data. Furthermore, scores in the comprehension section of Figure 3 (step 3.12;1) could not be accrued unless he asked a factual question. It is not surprising, therefore, that the scoring routine in the simulation failed to elicit independent responses for comprehension and factual behaviors.

The so-called reasonable convergent validity estimates for comprehension ($r_{14,11} = .40$; $r_{17,11} = .47$; and $r_{17,14} = .44$) in Table 1 have not been consistently discriminatory. The following is a similar discussion for those "reasonable" convergent validity estimates in problem solving subtest--namely $r_{15,12} = .49$ and $r_{18,12} = .50$ in Table 1, submatrix D.

As shown below these two convergent validity estimates of interest for problem solving exceed all corresponding correlations in their heteromethod-heterotrait triangles.

Problem-Solving Convergent Values	>	Corresponding Heteromethod-heterotrait Values
$r_{15,12} = .49$	>	$r_{13,12} = .26$; $r_{14,12} = .41$; $r_{15,10} = .39$; $r_{15,11} = .29$
$r_{18,12} = .50$	>	$r_{16,12} = .43$; $r_{17,12} = .48$; $r_{18,10} = .22$; $r_{18,11} = .46$

In addition to the above, these convergent validity estimates also exceed some (but not all) corresponding values in the mono-method-heterotrait triangles.

Problem Solving Convergent Values		Corresponding Monomethod-heterotrait Values
$r_{15,12} = .49$	>	$r_{12,10} = .42$; $r_{15,13} = .08$; $r_{15,14} = .37$
$r_{15,12} = .49$	<	$r_{12,11} = .54$
$r_{18,12} = .50$	>	$r_{12,10} = .42$
$r_{18,12} = .50$	<	$r_{12,11} = .54$; $r_{18,16} = .82$; $r_{18,17} = .88$

Other problem solving validity estimates that were significantly different from zero (but regarded as too low to warrant further inspection) do display some evidence of discriminate validity, in particular, those correlations involving simulation problem solving subtests. For example, in Table 1, $r_{6,3}=.36$ (correlation between simulation and multiple choice problem solving subtests) exceeds three out of four corresponding values ($r_{4,3}=.23$; $r_{5,3}=.02$; and $r_{6,2}=.32$) in the heterotrait-heteromethod triangles, and also exceeds both corresponding values in the simulation monomethod triangle ($r_{6,4}=.11$ and $r_{6,5}=.07$). Similar evidence of some discriminate validity is evident in the following correlations involving problem solving subtests with simulation, including those estimates in which time was confounded.

Problem Solving Convergent Values	Subtests Correlated
$r_{18,15} = .30$	S_{1p} E_{2p}
$r_{12,6} = .34$	S_{2p} M_{2p}
$r_{15,9} = .37$	S_{1p} E_{1p}
$r_{18,6} = .28$	S_{2p} E_{2p}

Since the absolute value of these is relatively low, one must infer that little validity is present. It is interesting to note, however, that

simulation appears to have elicited some behavior characteristic of what this study has called problem solving.

The last criterion for determining discriminate validity (according to Campbell and Fiske) is the relationship of traits within both heterotrait-heteromethod triangles and heterotrait-monomethod triangles. Inspection of Table 1 indicates that inter-trait relationships have not been consistent. For example, within the monomethod (solid) triangles for the first multiple choice form (submatrix A) and the second multiple choice form (submatrix D) problem solving and comprehension traits show the highest intercorrelations. Monomethod values for essay generally agree with this, but for simulation the highest correlations are observed between factual and comprehension traits; (previous discussion dealt with possible explanations for this latter case (cf. pp. 93-94). With the exception of $r_{3,1}=.43$ the one consistency that is evident in all 12 monomethod triangles is that factual-problem solving behaviors are correlated the least. This latter consistency, however, is not observed by certain correlations in the heterotrait-heteromethod triangles. For example, all correlations between simulation problem-solving and multiple choice factual ($r_{6,1}=.44$; $r_{15,10}=.39$; $r_{10,6}=.34$; $r_{15,1}=.43$) are consistently the highest correlations in their respective heterotrait-heteromethod triangles. Before inspecting these, however, the reader should note that the above four estimates are the correlations between simulation problem solving subtest (S_P) with the multiple choice factual (M_F) subtest. As noted these values are the highest in each of their own heterotrait-heteromethod triangles. However, the correlations between the simulation factual subtests (S_F) are not always highly correlated with the multiple problem solving subtests (M_P) (see $r_{4,3}=.23$; $r_{13,12}=.26$; $r_{12,4}=.26$; $r_{13,3}=.19$).

The fact that consistency was not observed between the intertrait relationships of S_P-M_F and S_F-M_P leads one to believe that response mode due to method (not response mode due to familiarity of data) accounts for some of the variance. Simulation problem solving scores were obtained by the candidate selecting or not selecting statements that were provided for him (cf. pp. 67-68). The selection was essentially a true-false selection. It is possible that some selection behavior of the correct alternative in multiple choice exams is very similar to this type of behavior in simulation problem solving--hence, the high intercorrelations of all subtests involving S_P and M_F. However, factual behavior in the simulation programs required the candidate to select one or more statements from a relatively long list. After each behavioral response, he was given information in visual or audio form (cf. Figure 3, step 3.12;1) that he had to comprehend. Following this he was asked questions regarding pertinency (step 3.12;2) and possibly even an additional question (step 3.12;3). Most or all the above took place before the candidate was returned to another simulation factual question. Therefore, the response mode for factual scores in simulation (S_F) was not as similar to the response mode of any multiple choice subtest, as were those responses in S_P. Hence the discrepancy between S_P-M_F and S_F-M_P correlations can be intuitively explained.

Relationships between traits involving the essays (E_1 and E_2) and the simulations (S_1 and S_2) are quite consistent, with comprehension and problem solving generally showing the highest intercorrelations, followed by factual-comprehension intercorrelations.

The heteromethod blocks for multiple choice and simulation follow similar patterns except that factual comprehension intercorrelations are usually the highest, followed by comprehension-problem solving correlations. In both the E-S and E-M heteromethod blocks, however, factual-problem solving relationships are usually the lowest. This is a small degree of evidence that behaviors were being elicited in a manner originally hypothesized (cf. p. 36).

For summary purposes, the reader should note that the following information was evident in the multitrait-multimethod matrix. (1) Serious limitations were encountered in estimating factual, comprehension, and problem solving validity because reliability amongst alternate forms was not good. (2) Many of the validity estimates for factual and comprehension variables were probably not statistically significantly different from zero. (3) Of those that were significant, many did not warrant further inspection because of their limited size. (4) Only variables involving simulation problem solving scores demonstrated any consistent, reasonable degree of discriminate validity. (5) Intertrait patterns within monomethod blocks indicate method variance was possible, particularly in the essays, and least in the simulations. (6) Trait patterns within monomethod blocks indicated that factual and problem solving scores were the least related traits. (7) Similar to the conclusion of point (6), trait patterns with heteromethod blocks indicated that factual and problem solving scores were least related. Some exceptions were noted, however, particularly for S_P-M_F; in this case it was assumed correlational patterns were due to similarities of response mode

in each of the methods. (8) Pattern consistency between factual-comprehension and comprehension-problem solving was not evident; depending upon which method (or methods) involved, either of these pairs demonstrated the highest interrelationship. In most cases, however, they ordered their intercorrelational magnitudes higher than factual-problem solving correlations.

The above inspection was essentially a conceptual appraisal. The influence of method variance and covariance as opposed to trait variance and covariance was difficult to estimate. Consequently, validity values of dubious size had to be ignored since irrelevant method or irrelevant intertrait influence was a possible source. Furthermore, establishing discriminant validity was difficult since certain convergent validity values did display some discriminating characteristic, but not always consistently. This was particularly true when validity values were being compared with monomethod values. To help determine the degree of trait versus degree of method influence in the multitrait-multimethod matrices, we now turn our attention to the factor analytic solution via restricted maximum likelihood estimates.

Results of Factor Analytic Solution

It was originally hypothesized that the simultaneous analysis of the 18 x 18 correlational matrix would yield a general factor, 3 method factors, and 3 trait factors. The manner in which factor loadings and factor correlations were restricted and allowed to vary for this original model was given in Figures 2A and 2B. The results corresponding to that model are provided in Table 2. Included in this table is a vector of specific

TABLE 2

RESTRICTED MAXIMUM LIKELIHOOD
(INCOMPLETE SOLUTION FOR ORIGINAL MODEL)

	General Factor	Three Method Factors			Three Trait Factors			
		M.C.	Sim.	Ess.	Fact.	Comp.	Prob.	
M_{1f}	-0.338	0.966	0.0	0.0	0.441	0.0	0.0	0.648
M_{1c}	-1.147	0.753	0.0	0.0	0.0	0.214	0.0	0.747
M_{1p}	-0.699	0.204	0.0	0.0	0.0	0.0	1.471	0.806
M_{2f}	-0.636	0.868	0.0	0.0	0.307	0.0	0.0	0.766
M_{2c}	-1.372	0.701	0.0	0.0	0.0	0.221	0.0	0.724
M_{2p}	-1.249	0.106	0.0	0.0	0.0	0.0	1.836	0.681
S_{1f}	-0.387	0.0	0.589	0.0	1.301	0.0	0.0	0.406
S_{1c}	-1.096	0.0	0.786	0.0	0.0	1.173	0.0	-0.442
S_{1p}	-0.415	0.0	-0.010	0.0	0.0	0.0	2.024	0.746
S_{2f}	-0.972	0.0	0.335	0.0	0.714	0.0	0.0	0.808
S_{2c}	-0.192	0.0	0.459	0.0	0.0	0.752	0.0	0.857
S_{2p}	-0.276	0.0	0.003	0.0	0.0	0.0	1.882	0.797
E_{1f}	-0.578	0.0	0.0	2.207	0.111	0.0	0.0	0.483
E_{1c}	-0.235	0.0	0.0	2.122	0.0	0.160	0.0	0.021
E_{1p}	-0.309	0.0	0.0	2.724	0.0	0.0	0.055	0.494
E_{2f}	-2.741	0.0	0.0	1.011	0.017	0.0	0.0	0.369
E_{2c}	-3.003	0.0	0.0	1.033	0.0	-0.011	0.0	0.094
E_{2p}	-2.563	0.0	0.0	0.724	0.0	0.0	0.615	0.445

[L] [E]

	G	M	S	E	F	C	P
G	1.0						
M	0.0	1.0					
S	0.0	-0.934	1.0				
E	0.0	-0.145	-1.118	1.0			
F	0.0	0.534	0.419	0.545	1.0		
C	0.0	0.745	0.282	0.834	-0.050	1.0	
P	0.0	0.114	-0.268	0.059	0.181	0.299	1.0

[P]

χ^2 with 84 df, is 78.4971
probability level is 0.649

error loadings for each of the 18 subtest scores. As described previously (cf. p. 44) this diagonal matrix was used in the denominator for estimating the probability that the model was an adequate fit. It also provides an estimate of the degree of extraneous variance which must be regarded as error.

The results of Table 9 are not meaningful, and should be regarded only as evidence that the solution attempted did not yield factors originally hypothesized. Also the solution is not complete, even though the IBM 360/67 computer made 183 iterations (over a time period of more than 25 minutes) attempting to minimize the function G. Since factor loadings in Table 9 were already beyond reasonable boundary limits, function minimization for the original model was not deemed to be necessary. (The results of Table 9 correspond to a minimization differences of .00012 for the last two iterations.)

By rejecting the model six different times (according to previously described criteria (cf. pp. 43-44)) a reasonable, interpretable solution was obtained. Table 3 in this chapter provides the results of this solution. The model for Table 3 corresponds to a general factor, an essay method factor, a collapsed factor involving factual and comprehension scores, and finally, a problem solving factor.

The need for having collapsed factual and comprehension traits into one factor is not surprising when one considers the previous discussions of the multitrait-multimethod matrices: (1) many of the factual and comprehension convergent validity estimates were relatively small; (2) factual-comprehension intercorrelations in all monomethod triangles were reasonably high; (3) most factual-comprehension intercorrelations in all

TABLE 3

RESTRICTED MAXIMUM LIKELIHOOD: SOLUTION I

	General Factor	Essay Factor	Fact and Comp Factor	Prob Factor	
M_{1F}	0.123	0.0	0.728	0.0	0.674
M_{1C}	0.362	0.0	0.442	0.0	0.821
M_{1P}	0.213	0.0	0.0	0.542	0.813
M_{2F}	0.183	0.0	0.536	0.0	0.824
M_{2C}	0.451	0.0	0.529	0.0	0.719
M_{2P}	0.412	0.0	0.0	0.625	0.664
S_{1F}	0.152	0.0	0.454	0.0	0.878
S_{1C}	0.385	0.0	0.511	0.0	0.769
S_{1P}	0.145	0.0	0.0	0.636	0.758
S_{2F}	0.334	0.0	0.348	0.0	0.876
S_{2C}	0.093	0.0	0.413	0.0	0.906
S_{2P}	0.106	0.0	0.0	0.577	0.810
E_{1F}	0.143	.903	-0.066	0.0	0.461
E_{1C}	0.057	1.018	-0.045	0.0	-0.000
E_{1P}	0.026	.932	0.0	0.004	0.356
E_{2F}	0.852	.418	-0.110	0.0	0.357
E_{2C}	0.926	.380	-0.066	0.0	0.128
E_{2P}	0.810	.350	0.0	0.136	0.434

[L] [E]

	G	E	F-C	P
G	1.000			
E	0.0	1.000		
F-C	0.0	0.450	1.000	
P	0.0	0.520	0.841	1.000

[P]

χ^2 with 108 degrees of freedom is 131.9095
(probability level is 0.059)

heteromethod triangles were also high, and often the highest (or at least the second highest) inter-trait relationship observed. The evidence in Table 3 is therefore a succinct description of what was previously known.

It is possible that defining factual behavior as that which conforms to instruction and experience incorporates definite weaknesses. It is to be remembered (cf. p. 8) that comprehension as defined by Bloom et al (1956) was very possibly included in this study's definition of factual behavior. Indeed some of the questions in various subtests were assumed to do this very thing (see discussion of scoring factual behaviors in the simulation programs, pp. 63-66). It appears, however, that the specialists in this study who tried to estimate information which was familiar due to experience (as well as instruction) did not successfully delineate such information from that which was regarded as novel (i.e., comprehension as defined in this study). Two possibilities (at least) may account for this. (1) Both factual and comprehension subtests measured only factual behavior. (2) Both factual and comprehension subtests measured only comprehension behavior.

The former possibility appears to be the least reasonable explanation. In previous discussions in this chapter (cf. p. 97) reasons were cited why simulation subtests for factual behavior probably required more than minimal mental effort. Furthermore, many of the multiple choice items classified as factual probably were not supported

by instructional procedures; this would be required by Bloom in order for them to be regarded as knowledge (i.e., factual). (Since the essay subtests did not load on the third factor of Table 3, intuitive inspection of this method adds little to the description of the factual-comprehension factor.) If one chooses to invoke the above reasons for rejecting the factual-comprehension trait as describing factual behavior, the possibility that it lies in a space described as comprehension appears to be reasonable.

The reader will notice that the first essay for comprehension does not meet the criteria of an adequate solution in that its loading on the essay factor exceeds unity, and accordingly the specific error loading for this variable is not greater than zero. However, Boruch and Wolins (1970), and Boruch, Larkin, Wolins and MacKinney (1970) maintain that solutions yielding a factor loading which does not exceed 1.02 may be regarded as acceptable within rounding error. This study is assuming this to be the case.

We now turn our attention to a more detailed inspection of Table 3. Squaring and summing the elements of E one observes that approximately 46% of test variance in the total matrix is regarded as error variance. This is not in conflict with the tentative conclusions at which one arrives when inspecting the reliability coefficients previously discussed in Table 1. Accordingly, the loadings in L account for only 54% of the observed variance.

The first column of L, the general ability factor, was constrained to be orthogonal to the remaining three oblique factors; (see matrix P for

intercorrelation among factors). Since all subtests were allowed to load on this factor it is assumed that whatever behavioral reactions that were common to all tests--probably general competence, particularly in the content area of applied statistics--loaded on this factor. Squaring and summing the eighteen elements of this factor indicates that approximately 17-1/2% of the total test variance, or 31% of the accountable variance is described by this general factor. The high loadings of essay 2 essentially define this factor. However, by inspecting the type of questions asked in essay 2 (cf. Appendix B.2) it is not unreasonable to assume that this factor is still a general ability factor, not particularly one that is characteristic of essays. The content assessed in essay 2 is a general problem often encountered in applied statistics courses, dealing with the relationship of sample size, level of significance, power of test, etc. Furthermore, the factor called essay method has high loadings for E_1 and reasonably high positive loadings for E_2. Accordingly, this latter factor most likely accounts for variance of the essay method. The correlations of the essay factor with the remaining two trait factors are further support of this interpretation since it is generally regarded by advocates of essays that this method lends itself to measuring behavior that this study called problem solving, and less so to the measurement of lower cognitive skills. As seen in the matrix P, the correlation of essay factor and problem solving is .52, and less (.45) with the factual-comprehension factor.

However, the respective essay trait subtests do not load, to any reasonable degree, on either of the two trait factors. For this solution

one may conclude that the two essays (marked in the manner previously described (cf. pp.51-53)) did not elicit behavioral patterns corresponding to factual-comprehension and problem solving traits.

This solution did not need to include any method factors for simulation and multiple choice subtests. To interpret the trait loadings for these methods, the relative size of loadings on the general factor and trait factor should be taken into account. If the loadings are comparable, the particular subtest did not elicit trait behavior any more than it elicited a general ability. For example, M_{1C} has a loading of .362 on the general factor and .442 on the factual comprehension factor; consequently little can be said about this subtest's unique ability in measuring behavior described as a combination of factual and comprehension.

On the other hand if column loadings in any row are relatively different, with the highest loading being on the trait, one can assume some degree of validity was obtained for a particular subtest. For example, S_{1P} has a general factor loading of .145 and a problem solving loading of .636. Accordingly, approximately 40.5% ($.636^2$) of this subtest's total variance, or 94% ($0.636^2 \div (1-.758^2)$) of this subtest non-error variance is accounted for by its trait loading. Since S_{2P} also has a relatively high loading on its trait factor (as compared to its general factor loading), results via factor analysis are in agreement with those conclusions made in Table 1 for the multitrait-multimethod matrix. That is, the simulation subtests for problem solving did have some validity.

The problem solving loadings for the multiple choice subtests, while comparable in size to the simulation loadings, are not as clearly dis-

tinguishable from their general factor loadings. Even M_{1P} which has the higher separation of the two, elicits some general ability almost as well as problem solving behavior.

In the first trait factor, called factual-comprehension, it is evident that M_{1F}, M_{2F}, S_{1F} and S_{2C} subtests load reasonably well on this factor and are discriminatory relative to their loadings on the general factor. On the other hand M_{1C}, M_{2C}, S_{1C}, and S_{2F} have loadings that are not discriminatory from corresponding weights on the general factor. The inconsistencies observed are not in conflict with the previous discussions of trait validity for these subtests. (As in the case of the problem solving factor, the essays did not load on this factual-comprehension factor.)

Since the first eight loadings which define the factual-comprehension trait factor, required subjects to select rather than compose answers, the possibility exists that the factor is actually not a trait, but a method or response mode factor. Attempts to find an adequate solution incorporating both methods and traits for multiple choice and simulation were fruitless. One of these attempts included a collapsed method factor defined as selected answer response; (i.e., all the multiple choice and simulation subtests were allowed to load on one factor). However, the only other reasonable solution found was that reported in Table 4. The model for this solution hypothesized a general factor and three method factors corresponding to the multiple choice, simulation, and essay formats. The first factor was constrained to be orthogonal to three remaining oblique factors. As seen in P the multiple choice and simulation factors

TABLE 4

RESTRICTED MAXIMUM LIKELIHOOD: SOLUTION II

	General Factor	Multiple Choice Factor	Simulation Factor	Essay Factor	
M_{1F}	0.296	0.593	0.0	0.0	0.749
M_{1C}	0.388	0.547	0.0	0.0	0.741
M_{1P}	0.258	0.551	0.0	0.0	0.794
M_{2F}	0.211	0.627	0.0	0.0	0.750
M_{2C}	0.466	0.549	0.0	0.0	0.693
M_{2P}	0.503	0.504	0.0	0.0	0.703
S_{1F}	0.140	0.0	0.773	0.0	0.619
S_{1C}	0.465	0.0	0.678	0.0	0.569
S_{1P}	0.330	0.0	0.201	0.0	0.922
S_{2F}	0.275	0.0	0.496	0.0	0.824
S_{2C}	0.209	0.0	0.483	0.0	0.850
S_{2P}	0.286	0.0	0.094	0.0	0.954
E_{1F}	0.701	0.0	0.0	0.543	0.462
E_{1C}	0.724	0.0	0.0	0.690	-0.000
E_{1P}	0.656	0.0	0.0	0.665	0.358
E_{2F}	0.875	0.0	0.0	-0.309	0.373
E_{2C}	0.922	0.0	0.0	-0.381	-0.073
E_{2P}	0.839	0.0	0.0	-0.275	0.470

[L] [E]

[P]

	G	M	S	E
G	1.000			
M	0.0	1.000		
S	0.0	0.430	1.000	
E	0.0	-0.017	-0.035	1.000

χ^2 with 114 degrees of freedom is 129.0103
(probability level is 0.159)

correlate 0.43, both of which are almost orthogonal to the essay method (-0.017 and -0.035 respectively). Like the results in Table 3, the solution does violate theoretical requirements since the specific errors for E_{1C} and E_{2C} are less than zero. Results for E_{1C} were violated in Table 3; but in both Tables 3 and 4 the solution for E_{1C} seems reasonable within rounding error. The case for E_{2C} in Table 4 is of a more serious nature since more than 105% of the variance is supposedly accounted for--an impossible situation. This author has still chosen to include these results, however, not so much as an accurate description of the data, but more as a verification of the interpretations made to the solution of Table 3.

The proportion of specific error variance is somewhat less in Table 4 --approximately 45%. This of course, is reflected in the differences of the observed probabilities of the χ^2's in Table 3 and Table 4. The proportion of variance accounted for in the general factor in Table 4 is approximately 28.5%. Not only is this variance proportion larger than for the corresponding factor in Table 3, but also the respective loadings indicate that both essays now load quite heavily on it. While these latter loadings greatly determine the definition of the factor space, it can be assumed general ability is still part of its destinctive characteristic since M_{1C}, M_{2C}, M_{2P}, and S_{1C} make reasonable contributions.

The remaining three method factors account for 31.7%, 26.4%, and 25.5% of the variance for all multiple choice, simulation and essay factors respectively. Except for M_{2F}, the multiple choice factor indicates that all subtests within this format elicit almost as much general ability

behavior as behavior related specifically to the method. In the case of the simulation factor, S_{1P} and particularly S_{2P} do not load heavily on these factors. If general simulation responses were mostly due to method there is evidence that the problem solving subtests in simulation do not conform to the remaining four subtests within that method. Indirectly, this is supporting evidence for the conclusions drawn in Table 3--namely, the highest trait validity estimates were via problem solving simulation subtests. It is to be noted, however, that no adequate solution was discovered which incorporated both trait and method, and therefore the respective loadings of S_{1P} and S_{2P} simultaneously taken on both trait and method is not reportable.

Interpretations of the essay factor is most difficult. If the factor is regarded as a measure of method, loadings for both E_1 and E_2 should be positive. Since they are not, it is more reasonable to assume that content differences between the two essays accounts for the essay loadings in Table 4's last factor. (We have already discussed why this difference might have occurred in the general factor of Table 3.)

To repeat a point, however, the results in Table 4 must be interpreted very cautiously, considering the violations observed in E_{2C}. These results do conform to some of the observations in Tables 1--the multitrait-multimethod matrix. The monomethod blocks of this latter table indicate that method variance is a plausible explanation, particularly for the essay. Furthermore, method variance was considered the least plausible explanation for S_{1P} and S_{2P}, and this is evident in

the solution of Table 4 , and confirmed in the problem solving trait factor of Table 3 . Excluding essays, the majority of trait loadings in Table 3 do not cleanly discriminate themselves from the general factor. As discussed above, method loadings in Table 4 (in general) do not discriminate themselves from the general factor either. Taken in an appropriate cautionary manner, Table 4 is confirmatory evidence that validity estimates were poor for the majority of collected data.

Even if one chose to restrict his inspection to the results of Table 3, it is quite evident that the behavioral reactions to the administered subtests has been essentially explained by two dimensions. The first must be called a general ability factor, which some essays seem to measure quite well. The second dimension is in the area of cognitive skill as well, and which can be partially explained by the traits defined in this study. These traits, however, have not clearly delineated themselves. Factual and comprehension traits were unitary and the problem solving trait was found to be highly correlated (.84) with the factual-comprehension factor. With some exceptions, test behaviors indicate that responses due to a general ability are as great as the proportion of behavior that is elicited by the defined traits. These exceptions fall into three groups. Firstly, the essay does not elicit trait behavior as defined and measured in this study. Secondly, simulation appears to measure problem solving skills to some degree. Thirdly, multiple choice tests appear to measure a trait called factual-comprehension to some degree. For this latter case, however, responses due to the format of the multiple choice exams remains a possibility.

An alpha coefficient was calculated for the eighteen subtest scores and was found to be .818. If the number of units entering this composite were doubled (36 subtests) application of the Spearman-Brown Prophecy formula (Gulliksen, 1951, p. 78) indicates the coefficient would be as high as .90. Under Cronbach's interpretation that α reflects the degree of unifactoredness, these coefficients may indicate that measured behavior was that due to content--namely, the area of applied statistics and research design. Remembering that Table 3's results essentially described two orthogonal dimensions, it is not surprising to find a high α coefficient.

Since the possibility existed that individual subtests were poorly constructed, and thus limited the success of these findings, some additional analyses were done to estimate the degree of consistencies of components entering some subtests.

Multiple Choice Item Analyses

Item analyses done in this study were restricted to the multiple choice tests, and the results are reported in Tables 5 - 7. Each of these tables respectively deals with those items pre-classified as factual, comprehension, and problem solving. The top half of each table deals with trait items in the first multiple choice test (M_1), the bottom half of the tables with trait items in the second multiple choice test (M_2). The first two columns indicate the item number and code classification for each item; these match the item numbers and codes in Appendix A. The third, fourth, and fifth columns are reported biserial coefficients respectively

TABLE 5

ITEM ANALYSIS FOR MULTIPLE CHOICE: FACTUAL

	Item	Code	Biserial Coefficient			Difficulty (p)
			$M_{1:F}$	$M_{1+2:F}$	M_1^*	
Column	(1)	(2)	(3)	(4)	(5)	(6)
M_1	2	F1	.29	.18	.15	.84
	3	F2	.49	.36	.43	.84
	4	F3	.28	.39	.33	.44
	10	F4	.32	.25	.18	.24
	11	F5	.46	.24	.53	.30
	16	F6	.42	.34	.25	.72
	22	F7	.48	.38	.40	.64
	25	F8	.55	.43	.55	.60
	26	F9	.69	.42	.44	.68
	32	F10	.00	.00	.00	1.00
	36	F11	.54	.56	.31	.52
	38	F12	.53	.51	.33	.34
	39	F13	.47	.37	.37	.70
	41	F14	.42	.28	.05	.70
	44	F15	.40	.59	.53	..80

	Item	Code	$M_{2:F}$	$M_{1+2:F}$	M_2^{**}	Difficulty (p)
M_2	1	F1	.29	.13	.29	.56
	2	F2	.29	.16	.21	.82
	3	F3	.36	.06	.31	.36
	4	F4	.42	.49	.05	.76
	5	F5	.46	.54	.38	.32
	6	F6	.44	.37	.58	.94
	10	F7	.35	.47	.14	.92
	11	F8	.75	.59	.50	.54
	13	F9	.11	.05	.29	.92
	20	F10	.13	.23	.23	.88
	22	F11	.28	.28	.31	.60
	38	F12	.54	.41	.48	.60
	39	F13	.49	.16	.31	.34
	43	F14	.50	.43	.40	.31

* M_1: total multiple choice test, form 1

** M_2: total multiple choice test, form 2

TABLE 6

ITEM ANALYSIS FOR MULTIPLE CHOICE: COMPREHENSION

	Item	Code	Biserial Coefficient $M_{1:c}$	$M_{1+2:c}$	M_1^*	Difficulty (p)
Column	(1)	(])	(3)	(4)	(5)	(6)
M_1	1	C1	.67	.54	.60	.90
	5	C2	.31	.28	.44	.60
	8	C3	.14	.16	.28	.94
	9	C4	.29	.27	.26	.34
	12	C5	.21	.24	.28	.50
	13	C6	.51	.44	.28	.48
	17	C7	.29	.39	.14	.80
	18	C8	.54	.55	.55	.44
	19	C9	.62	.50	.42	.72
	23	C10	.38	.26	.27	.64
	24	C11	.33	.32	.46	.18
	29	C12	.46	.40	.33	.52
	30	C13	.23	.10	.34	.74
	31	C14	.55	.44	.34	.34
	33	C15	.48	.42	.45	.66
	35	C16	.75	.69	.50	.66
	37	C17	.61	.58	.37	.88
	40	C18	.20	.00	.14	.36

	Item	Code	$M_{2:c}$	$M_{1+2:c}$	M_2^{**}	Difficulty (p)
M_2	7	C1	.71	.74	.76	.40
	8	C2	.68	.59	.73	.36
	15	C3	.36	.28	.29	.74
	16	C4	.32	.41	.40	.52
	19	C5	.10	.01	.08	.46
	21	C6	.28	.28	.17	.34
	26	C7	.48	.39	.44	.82
	27	C8	.37	.44	.39	.22
	29	C9	.21	.11	.28	.26
	30	C10	.05	.24	.07	.68
	31	C11	.53	.53	.39	.44
	32	C12	.18	-.10	.09	.50
	33	C13	.42	.20	.20	.20
	34	C14	.32	.18	.21	.66
	35	C15	.30	.14	.23	.68
	36	C16	.45	.44	.47	.66
	41	C17	.57	.55	.49	.70
	42	C18	.74	.66	.63	.33

* M_1: total multiple choice test, form 1

** M_2: total multiple choice test, form 2

TABLE 7

ITEM ANALYSIS FOR MULTIPLE CHOICE: PROBLEM SOLVING

	Item	Code	Biserial Coefficient $M_{1:p}$	Biserial Coefficient $M_{1+2:p}$	M1*	Difficulty (p)
Column	(1)	(2)	(3)	(4)	(5)	(6)
M_1	6	P1	.54	.26	.50	.56
	7	P2	.26	.26	.40	.84
	14	P3	.19	.08	.02	.54
	15	P4	.51	.82	.79	.18
	20	P5	.63	.48	.37	.72
	21	P6	.41	.40	.32	.28
	27	P7	.25	.19	.17	.58
	28	P8	.40	.13	.28	.72
	34	P9	.28	.06	.16	.26
	42	P10	.29	.30	.14	.20
	43	P11	.70	.68	.48	.42
	45	P12	.41	.14	.31	.84

	Item	Code	$M_{2:P}$	$M_{1+2:P}$	M2**	Difficulty (p)
M_2	9	P1	.33	.35	.40	.16
	12	P2	.50	.31	.35	.44
	14	P3	.21	.27	.12	.54
	17	P4	.82	.24	.86	.12
	18	P5	.77	.76	.46	.20
	23	P6	.58	.50	.54	.34
	24	P7	.37	.30	.05	.32
	25	P8	.75	.67	.67	.50
	28	P9	.66	.64	.64	.32
	37	P10	.37	.52	.26	.82
	40	P11	.72	.71	.90	.12
	44	P12	.55	.50	.37	.46
	45	P13	.43	.31	.38	.18

* M_1: total multiple choice test, form 1

** M_2: total multiple choice test, from 2

calculated on a unit test of: (1) the same trait items from one multiple choice test (column 3), (2) the same trait items from both multiple choice tests (column 4), and (3) all items from any one multiple choice test (column 5). For example, the third column in Table 5 reports the estimated correlation between factual items and a total score where the total score is either 15 factual items in M_1, or 14 factual items in M_2. Total score in column four is modified to include the combined score on all factual items in M_1 and M_2. The fifth column uses the total score obtained in either M_1 or M_2, both of which include comprehension and problem solving items. The sixth and last column of Tables 5 - 7 reports the proportion of candidates which answered each question correctly. (It is evident items were not consistent in their difficulty level.)

If one uses a criterion of 0.35 for determining if items have respectable biserial coefficients, a majority of the trait items meet or exceed this standard, regardless of the composition of the unit test. The use of items which had shown previous discriminatory power probably contributes to these results. (The five problem solving items written by this author also exceeded the criterion standard of 0.35: see Table 7--M_1 (P4, P12) and M_2 (P5, P12 and P13).)

For eliciting defined trait behavior, items must also meet additional criteria. It is assumed the total score in columns three and four describes one trait (e.g., factual in Table 5), and the fifth column reflects a total score in a domain of all traits. Consequently, an item which is purported to measure the one trait should correlate higher with the former two total scores than with a total score reflecting many traits.

For example, the item M_1: F_1 should correlate higher with the total score in $M_{1:F}$ (column 3) and $M_{1+2:F}$ (column 4) than with M_1 (column 5). Since alternate form reliability estimates are reflected in differences between column 3 and 4, one might expect correlations in the former to be highest.

The reader will notice that few biserial coefficients in Tables 5-7 exceed 0.35 and properly order their values (as explained above) over columns 3-5. For example, the first factual item in Table 5 has values ordered as expected, but does not exceed 0.35 in either of columns 3 or 4. In the same table, item F_6 in M_2 exceeds 0.35 but the highest correlation observed is in column 5; this indicates the item measures behavior which is more characteristic of all traits than simply a factual trait. Items F_{11} and F_{12} in M_1, on the other hand, meet all criteria. Since the majority of items do not exhibit characteristics like these last two items, one has supporting evidence to former conclusions regarding the validity of multiple choice tests. That is, individual items as well as total scores in multiple choice subtests, did not adequately elicit behavior which this study defined as its objective of measurement. This is not to say that the multiple choice format is invalid for any measurements, for indeed it has shown some discriminatory power. It does imply, however, that this study was unable to construct multiple choice tests which elicited intended behavior for the sample of subjects used.

Based on the information of Tables 5-7, this author chose the best 40 items from M_1 and similarly for M_2 and attempted to factor analyze them to see if they loaded together in three trait clusters. The principal axis solution yielded 17 factors with eigenvalues greater than one. With

the exception of a few problem solving and comprehension items which loaded together, the factors were not interpretable. Attempts to define the factors in terms of Bloom's classification scheme were also fruitless. Since a principal axis solution yields orthogonal factors, Joreskog's technique was also used. A model corresponding to three oblique trait factors and one orthogonal general factor did not account for a significant proportion of the variance. Consequently, the need for more factors was indicated. Due to the fact that: (1) analyses up to this point had indicated the items had not elicited intended behaviors, and (2) the cost of further analyses (particularly for factor analysis) was high, no further efforts were made to determine in what other ways the items could be grouped.

The KR_{20} coefficients for all trait subtests, combined trait subtests, and whole tests are reported in Table 8. Since length of test affects this estimate, the Spearman-Brown Prophecy formula was applied to these coefficients. The last three columns of Table 8 report the modified coefficients if the tests had been of various unit lengths. Inspection of these last three columns indicates that the multiple choice items do have some reliability, when consistency is defined as inter-item agreement. (Possible exceptions to this statement are evident in M_{2F} and M_{1P}.) Since most coefficients are near .80, both for trait and total tests of 90 items in length, there is some indication that the influence of one particular factor is greatest. Whether that factor is test wiseness, achievement in general, or achievement specifically in the content area, is unknown.

TABLE 8

INTER-ITEM CONSISTENCY ESTIMATES (KR_{20})

(Multiple Choice Tests)

Multiple Choice Test	No.of Items	KR_{20}	S-B Prophecy Formula if No. of Items is		
			15	45	90
M_1 (whole)	45	.662	.395	.662	.797
M_2 (whole)	45	.692	.428	.692	.818
M_{1+2} (whole)	90	.815	.423	.678	.815
M_1 (fact)	15	.388	.388	.655	.792
M_2 (fact)	14	.171	.181	.399	.570
M_{1+2} (fact)	29	.536	.374	.642	.782
M_1 (comp)	18	.434	.390	.657	.793
M_2 (comp)	18	.380	.338	.605	.754
M_{1+2} (comp)	36	.608	.392	.654	.795
M_1 (prob)	12	.096	.117	.285	.443
M_2 (prob)	13	.500	.556	.776	.874
M_{1+2} (prob)	25	.55	.423	.688	.815

In summary, the multiple choice tests were reasonably reliable and discrimatory when inspected by normal procedures using statistics such as the KR_{20} and biserial coefficients. However, more rigorous inspections indicated they were not measuring well the objectives this study had set out for them.

Inter-Judge Consistencies

The reader will recall that several controls were introduced into this study, in order that essay markers' inconsistencies would be reduced (cf. pp. 50-53). The estimated degree of consistency realized in this study is reported in Table 9. Since the final score a candidate obtained was the average of each marker's standardized z-score, the estimated consistency for any one judge is that reported in column 2 of Table 9. These values indicate the degree of linear relationship between the two markers for each essay subtest. The Pearson product-moment coefficient was calculated. If differences of marker variances had not been extracted in the z-scores, column 3 indicates the reliability of any one judge, where inconsistencies of linear relationship and variances are reflected in the reported values. Due to the fact that two scores more accurately estimate the true score of any individual, column 4 indicates the estimated inter-judge consistency when the average score is used. Accordingly the reader will see that column 4, has higher reliability estimates than column 3. In an analogous manner, columns 5 and 6 indicate the reliabilities of one and two judges respectively, but when mean differences are also regarded as a

TABLE 9

INTER-JUDGE CONSISTENCY ON ESSAYS

Col. 1	Col. 2	Col.3	Col.4	Col.5	Col.6	Col.7	Col.8	Col.9	Col.10	Col.11
Subtest	Linear Relationship	Consistency of Linear Relationship & Variances		Consistency of Linear Relationship, Variances & Mean		Judge 1		Judge 2		Probability for Differences of $\bar{X}$'s
						$\bar{X}$	SD	$\bar{X}$	SD	
E_{1F}	0.655	0.667	0.800	0.660	0.795	8.62	3.24	8.08	3.04	0.143
E_{1C}	0.692	0.705	0.827	0.701	0.824	6.08	3.39	5.56	3.22	0.188
E_{1P}	0.768	0.783	0.878	0.787	0.880	3.70	3.96	3.82	3.89	0.744
E_{2F}	0.696	0.702	0.825	0.403	0.575	10.02	3.26	7.20	2.79	0.000
E_{2C}	0.698	0.690	0.817	0.475	0.644	8.46	5.98	3.58	5.98	0.475
E_{2P}	0.519	0.501	0.667	0.492	0.660	5.78	3.96	5.08	2.81	0.155

source of inconsistency. Columns 7-10 of Table 9 indicate the mean and standard deviation of each judge's scores; column 11 provides a probability estimate of population mean differences.

The reader will see that if variances for the two judges are quite similar, estimates in column 3 and column 5 are close to being equivalent. In general, markers were not inconsistent in the amount of dispersion when assigning marks. The markers were not as consistent in the mean assignment, however, For example, E_{2F} subtest has means of 10.02 and 7.20 for judge 1 and 2 respectively. Consequently, the reliability estimate for this subtest (reflecting mean inconsistencies), is less than any estimate in which mean differences are not important. In summary, markers were not inconsistent in mean marks for all subtests in E_1 and also E_{2P}; inconsistencies of mean marks were evident in E_{2F} and E_{2C}.

In conclusion, experimental controls to reduce reader inconsistencies in essay markers were effective in reducing the variance of marks of the two judges and only partially effective in controlling mean differences. In terms of controlling linear inconsistencies (which includes inconsistencies of candidate-rankings) this study has not demonstrated that the procedures elevated reliabilities. That is to say, the estimates in column 2 are comparable to those discussed by Hazlett, Maguire, and Wilson (1969). In this latter study over 60 pairs of markers (none of whom had keys, typed copies, training sessions, typed anonymous answers, etc.) demonstrated a similar degree of agreement as that observed in

this experimental situation. This author must concur with Huddleston's (1954) conclusion, "...there is no convincing evidence...that these hopes [i.e., the use of experimental controls to enhance consistencies of essay markers] have come to fruition". It is to be remembered, however, that estimates for essay reliability for alternate forms (observed in Table 1) are quite comparable to those for other methods. Consequently, this author must admit he was unsuccessful in establishing any respectable reliability estimates for any method.

Mean Differences in Sample

Theoretically, subjects who have been more thoroughly trained in a specified content area should display higher degrees of trait behavior than subjects who have had less training. Consequently, mean differences among such groups should be observed in all subtests administered in this study. Twenty-seven of the subjects tested in this study were registered in courses at the time of data collection; they have been labelled group 1 and 2 (cf. p. 71). The remaining subjects had completed at least these courses and were labelled group 3. 't' tests were done to determine significant differences between means in hypothetical populations corresponding to groups 1 and 2 and group 3. Assuming group 3 should perform more adequately than groups 1 and 2, a one-tail t-test at the 0.05 level of significance was considered appropriate. Table 10 reports the results of these tests where means for group 3 were subtracted from those of groups 1 and 2. Welch's procedure (Ferguson, 1966, pp. 172-173) for controlling variance differences has been used in these tests.

TABLE 10

MEAN DIFFERENCES IN SAMPLE

(Welch T Prime Adjustment of T Tests)

Method-Trait Subtest	Group 1 & 2 (n=27) $\bar{x}$ (S.D.)	Group 3 (n=23) $\bar{x}$ (S.D.)	Adj. df.	t'	P-One Tail
M_{1F}	8.89 (2.10)	10.35 (1.90)	47.81	-2.579	0.0065*
M_{1C}	9.81 (2.51)	11.78 (2.09)	47.98	-3.025	0.0020*
M_{1P}	5.81 (1.44)	6.52 (1.78)	42.29	-1.525	0.0673
M_{2F}	7.93 (1.49)	9.96 (1.36)	47.73	-5.024	0.0000*
M_{2C}	8.15 (1.73)	9.91 (2.95)	34.21	-2.523	0.0082*
M_{2P}	4.11 (1.72)	5.00 (2.65)	36.63	-1.382	0.0876
S_{1F}	22.56 (5.38)	25.87 (4.67)	47.97	-2.331	0.0120*
S_{1C}	18.63 (5.56)	22.13 (5.53)	46.83	-2.225	0.0155*
S_{1P}	14.63 (1.94)	16.35 (4.25)	29.75	-1.786	0.0421*
S_{2F}	23.67 (5.78)	23.35 (6.29)	45.24	0.185	------
S_{2C}	24.70 (5.91)	24.65 (5.53)	47.54	0.032	------
S_{2P}	12.56 (2.61)	13.43 (4.15)	35.82	-0.878	0.1928
E_{1F}	-0.16 (0.99)	0.18 (0.80)	47.87	-1.345	0.0925
E_{1C}	-0.18 (0.98)	0.21 (0.83)	48.00	-1.495	0.0708
E_{1P}	-0.15 (0.97)	0.18 (0.90)	47.69	-1.238	0.1108
E_{2F}	-0.17 (0.98)	0.20 (0.83)	48.00	-1.447	0.0772
E_{2C}	-0.15 (0.91)	0.17 (0.93)	46.32	-1.220	0.1143
E_{2P}	-0.21 (0.81)	0.24 (0.91)	44.52	-1.824	0.0374*

*Significant at the .05 level

Except for S_{2F} and S_{2C} mean differences are in the expected direction, even though some are not statistically significant. It is important to note that these differences are not necessarily due to differences of defined trait behavior. It can only be assumed that some differences were observed in scores assumed to be indicative of defined trait behavior. Consequently, results of Table 10 contribute little to the establishment of construct validity.

Face Validity Estimates

Face validity is essentially a measure of that which subjects thought they did when reacting to each of the methods. Questions were posed to the subjects after all tests had been completed. Consequently, responses were not to individual subtests but to both forms of each method. While these responses do not establish an accurate description of candidate behavior, they do indicate if subjects' attitudes agree with responses made on various subtests. Furthermore, the responses to the first three questions (cf. Appendix D) provide the reader with information which indirectly indicates some possible sources of invalidity in previously established estimates of construct validity.

The measurement of opinions (like any other behavioral measurements) must be first reliable before it can be assumed to be accurate or valid. To estimate reliability of the questionnaire, two questions were asked for each measured opinion. Those pairs have been previously described (cf. p. 75). For each pair of items it was assumed that at least two out of every three respondents had to make a similar response to both

questions before any further inspection was made of the data contained in those two items. That is to say, unless 67% of the candidates consistently answered two questions (both of which tested for the same information) then it was assumed the answers were not worth inspection. Table 11 provides the reader with the percentage of agreement obtained for all pairs of items. In general, the consistency rate was good, with the exception of: (1) the pair of items dealing with the multiple choice measurement of Bloom's trait called analysis, (2) the two pairs of items dealing with the essay measurement of Bloom's traits called syntheses and evaluation, (3) the pair of items which asked the candidate to indicate if the information tested in the essays was relevant or irrelevant. These four questions do not exhibit consistency, and therefore provide a poor basis for reporting any further information. It is interesting to note that candidates' responses were not consistent when asked if they thought the essays had measured traits of syntheses and evaluation. Proponents of essays cite this method as the best vehicle for assessing these traits; a reasonable number of candidates writing the essays in this study thought otherwise.

For further inspection of consistent pairs of items in Table 11 we now turn our attention to the percentage of candidates who at least agreed in both items that a particular method did indeed measure a specified trait. For these results see Table 12. Note that Table 11 indicates the percentage of consistent responses to item pairs. Table 12 indicates the percentage of all 50 subjects who marked either agree or strongly agree to both items in any pair. Again a criterion of 67% is invoked before

TABLE 11

RESPONSE CONSISTENCY TO QUESTIONNAIRE
(% of Agreement on Specified Pairs of Items)

	Multiple Choice	Simulation	Essay
Factual	86*	78*	74*
Comprehension			
Interpretation	72*	76*	78*
Application	68*	88*	74*
Analysis	64	86*	78*
Problem Solving			
Synthesis	70*	80*	44
Evaluation	70*	80*	58
Relevant	68*	86*	64

*Deserves further inspection.

TABLE 12

FACE VALIDITY ESTIMATES FOR CONSISTENT ITEMS
(% of subjects indicating method measured trait)

	Multiple Choice	Simulation	Essay
Factual	54	20	22
Comprehension			
Interpretation	38	74*	72*
Application	10	82*	46
Analysis	--	82*	58
Problem Solving			
Synthesis	4	80*	--
Evaluation	14	76*	--
Relevant	64	86*	--
Too Difficult?	%	%	%
Total group	24(yes)	48(yes)	36(yes)
Group 1 & 2	33(yes)	59(yes)	48(yes)
Group 3	13(yes)	35(yes)	22(yes)

*Method has face validity for the measurement of specified trait.

this author is willing to state that most candidates stated that a particular method measured a specified trait. According to this criterion the multiple choice tests did not have reasonable face validity for measuring any trait. The simulation programs had face validity for all traits except for factual behavior. The essay had face validity for measuring only what Bloom calls interpretation. Furthermore, only the simulation programs were consistently regarded as relevant to the measurement of applied statistics and research design. Included at the bottom of this table is the percentage of agreeing responses to this single statement: the examination was too difficult. The percentages of responses has been reported for the total group as well as for groups 1 and 2 and group 3. It is evident the subjects who had less training in the content area regarded the examinations as the most difficult. In general, all groups considered the simulation tests as the most difficult, and the multiple choice examinations the least difficult.

How does the information of Table 12 bear upon attempts to establish construct validity? Firstly, since both the multiple choice and essay examinations did not appear to be relevant, it is possible such a factor inhibited the exhibition of trait behavior. The possibility exists, therefore, that poor attainment of construct validity was due to perceived irrelevant tasks required in the multiple choice and essay formats. On the other hand, subjects felt that comprehension behavior (as defined in this study) was measured in the simulation programs. Previous analyses indicate that if it was, the scoring routines were not sensitive to such measurement. However, previous validity results for problem solving measurement in the simulations are in agreement with the opinions of subjects.

Caution must be used when interpreting face validity estimates for simulation. Since it was not only a novel technique, but also regarded as the most difficult, opinions of subjects may reflect these variables rather than a true statement of the capacity for simulation to measure either comprehension or problem solving behaviors. It is noteworthy, however, that a trait called synthesis by Bloom was considered to be a measured behavior. The measurement of syntheses is usually done by essays; however, the candidates of this study have indicated that simulation, not essay, is the proper means for its measurement.

Summarization of Statistical Analyses

Empirical estimates of construct validity have not been noteworthy for most subtests administered in this investigation. A number of possibilities have been suggested why this was the case. (1) The traits used in this study did not lend themselves to measurement. (2) Scoring procedures did not properly cluster the respective trait behaviors. (3) Preclassification of subtests did not conform to the defined traits. (4) Respectable alternate form reliability was not achieved. (5) Method variance interfered with trait exhibition. (6) Measurement elicited behavioral reactions to content not operations on content. (7) Components entering subtests did not conform to the trait classification. (8) The irrelevant appearance of some subtests inhibited intended behavior.

One exception to the above comments was consistency observed--namely, the possibility that simulation may have elicited behavior which this study defined as problem solving. The inspection of the multitrait-

multimethod matrix, the oblique factor solution of its correlations, and the opinion of tested subjects all indicate that this method may have elicited problem solving behavior. The opinion of subjects may have been biased in favor of this finding since simulations were regarded as difficult. Consequently behavior in all or most simulation subtests may have been "problem solving". Furthermore, inferences based on the intercorrelations and factor solutions of the multitrait-multimethod matrices must inevitably use all methods to define problem solving. Consequently the degree to which the essay and multiple choice subtests evaluated problem solving also helps to define what behavior the simulation problem solving subtests actually measured. Since essay contributions were negligible, the multiple choice subtests remain as the only other format to be used for comparative purposes. Chapter Two has already cited the widely held opinion that multiple choice items do not measure skills typical of problem solving behavior. The possibility exists, therefore, that the problem trait measured by simulation is more characteristic of traits defined as application or analysis by Bloom, rather than characteristic of evaluation or synthesis. In rebuttal to this hypothesis, this author reminds the reader that both specialists and subjects regarded the simulation programs as a valid means for eliciting problem solving behavior so often classified as synthesis and/or evaluation.

This summary also cites the possibility that some of the multiple choice and simulation subtests did elicit some behavior characterized as either factual or comprehension--the latter being the more likely. Due to the fact that (1) these validity estimates were not consistent in all subtests, (2) these estimates were often equivalent to a general ability

trait, and (3) these estimates could possibly be explained in terms of method influence, this author has chosen a position of neutrality. That is to say, multiple choice and simulation tests may elicit factual or comprehension behavior, but the results of this study do not confirm such a statement.

Finally, it is reasonable to state that essays do measure some behavior which is common to all methods used in this investigation. Such behavior may well be one of many things: (1) test wiseness, (2) general achievement, (3) competence in applied statistics and research design, or (4) any other reasonable common denominator present at the time of this study's collection of data. There is additional evidence that behavior is elicited which is quite unique to essays. Proponents of the essay have some grounds for arguing that it is behavior which this study called problem solving. If this argument is valid, it goes without saying that the simulation programs were unable to tap this cognitive domain. It is the opinion of this author, however, that unique variance observed amongst the essays is more a function of the method or possibly of the markers. Neither essay method nor marker influence is considered to be <u>uniquely</u> suited to assessing problem solving behavior. If either is uniquely suited, achievement measurement in this area must be cautiously used because inter-judge reliabilities appear to have definite limits.

This last statement leads the reader into the area of what this study has contributed generally to the area of educational measurement and decision-making. Based on the findings of this chapter, these topics will now be discussed (cf. Chp. 5).

CHAPTER FIVE

SUMMARY AND RECOMMENDATIONS

Summary

In the introductory chapters of this report, the importance of reliable and valid measurements for achievement tests was emphasized, the latter being defined as the sufficient condition for adequate assessment. Reliability and validity were viewed in terms of one concept--consistency, (Campbell and Fiske, 1959). Reliability was seen as the consistency of maximally similar methods in measuring the same variables, and validity as the consistency of maximally different methods measuring those same variables. These variables--that is the basis on which consistency estimates are established--can in essence be any measure of interest to the evaluator. To be of use in the educational process such variables should meet the criteria set for behavioral objectives. These include characteristics such as (1) completeness, (2) consistency, (3) explicitness, and (4) relevancy. Additional standards such as the description of inferred and/or observable behavior, as well as directed and/or terminable behavior were also cited.

This study chose to use a modified classification of a cognitive hierarchy described by Bloom, et al (1956) as its statement of behavioral

objectives in graduate achievement examinations. This classification was defined by a test committee in the Royal College of Physicians and Surgeons of Canada. Like Bloom's description, the Committee assumed cognitive ability was composed of hierarchical levels. The lowest level of interest for measurement purposes was labelled factual, and described "minimal mental effort". In a departure from Bloom's corresponding description of knowledge, factual behavior was assumed to take place whenever familiar data were being tested--where teaching and/or experience was responsible for the data being regarded as familiar. Therefore some subcomponents of Bloom's next hierarchial level--comprehension--possibly was now being labelled as factual in this study.

The second variable of interest in this investigation was labelled comprehension. It is not to be confused with the identical label in Bloom's scheme. Comprehension in this study refers mostly to what Bloom describes as application and analysis, but also may include behaviors that he describes as interpretation and extrapolation. Such behavior was not considered to be a minimal mental exercise, and was assumed to take place whenever subjects dealt with moderately novel data.

The last variable defined in this modified scheme was labelled problem solving and was characterized by terms such as synthesis and evaluation (Bloom, 1956, pp. 162-197). For such behavior to be exhibited complex mental effort and very novel data were necessary conditions. Judgement, versatility, organizational and strategic ability were characteristic behaviors. By definition essay writing was not considered to be a necessary and/or sufficient demonstration of this trait.

This is not to say that such a skill is not problem solving; it does dictate that problem solving behavior is not necessarily described by essay writing.

These three traits were regarded as reasonable objectives in achievement measures. Consequently attempts to establish validity estimates were attempts to establish construct validity estimates. While other concepts of validity such as content, face, concurrent, and predictive validity are relevant, construct validity provides the most potential and generalized basis for educational decision making. This is not to say this study did not concern itself with establishing content, face, and concurrent validity. Indeed, certain data collected in this investigation were regarded as estimates of these three types of validity estimates. Their generalizability is minimal, however, and consequently little emphasis is placed in their results other than their contributions to the establishment of construct validity.

Since relatively independent methods were needed to estimate the existence of constructs, this study used two common, but seemingly independent, evaluation tools--the essay and the multiple choice tests. Research reviewed indicated that the essay has inherent face validity for measuring problem solving behavior. Due to the fact that essays can sample only a minimal amount of behavior, and also because inter-marker agreement is usually low, this technique has often been criticized by users of the multiple choice test. These latter users cite advantages such as reliability and broad sampling of behavior as the main reasons

for using multiple choice items. Criteria of worth must be in terms of validity, however, and this study rigorously maintains validity is the only criterion for determining a method's usefulness for evaluative purposes. A third method, the novel procedure of simulation, was also used in this investigation. This procedure is being used increasingly because, (1) it has the inherent face validity for measuring relevant behavior, particularly at the problem solving level, (2) it has the potential for sampling behavior as broadly as the multiple choice format, (3) it has the potential for developing standards of internal consistency equivalent to the multiple choice format, and (4) it is not hampered by the unstructured and unstandardized format of the essay which no doubt is at least partly responsible for poor consistencies observed among markers. Like the essay and multiple choice methods, however, the sufficient criterion for determining the evaluative worth of the simulation is in terms of its sensitivity in measuring those variables which the evaluator defines as his objectives for assessment.

To maximize the potential sensitivity and generalizability of each of the three methods in measuring factual, comprehension, and problem solving behavior, certain experimental design constraints were implemented in this investigation. For instance, 85 multiple choice items were selected from a pool of more than 500 available items. Selection was not random, a requirement for generalizing results of content validity to typically used multiple choice items. Instead selection was based on criteria such as clarity, suitability to taxonomic

classification, previous evaluator satisfaction including biserial and difficulty coefficients. Five additional problem solving items were also composed to insure an adequate sampling of this latter trait. These items were then classified according to this study's taxonomic scheme, and the unanimous agreement of three evaluators does provide some basis for generalizing the estimated degree of content validity for these items. That is to say, this consensus is a more accurate description of what evaluators in general would consider that the items measured rather than if the classification had been done by this researcher alone.

Similar use of specialists was incorporated for establishing the content validity of the essays and simulation problems. In the case of the essays, steps were also taken to possibly enhance intermarker agreement: (1) separate, typed copies of all answers were provided to each marker; (2) unique identification numbers replaced candidates names, (3) standard keys, including weighting schemes for scores were used; (4) six hours of marker training sessions acquainted the judges with the intent of the study; and (5) candidates' scores were the average of the two standardized scores assigned by the markers. In the case of the simulation problems, this report attempted to describe the manner in which this technique elicited trait behaviors. The extent to which the reader wishes to accept the simulation pathways as meaningful, will determine the degree of simulation's content validity. This study can report, however, that two specialists as well as this author assumed

respective scores obtained in this technique were indicative of factual, comprehension, and problem solving behavior and similar to the information tested in the multiple choice and essay tests.

Trait scores were obtained by administering two alternate forms of each method on two successive days. Reliability estimates, therefore, reflected consistencies of alternate forms over a period of one day. Nonrandom selections of graduate students were assigned to take each battery of tests in which the order within each battery was randomly determined. Tests were administered under non-stress conditions and anonymity of test performance was assured all candidates. These candidates were by inspection differently trained in the content area. Consequently directional mean differences were assumed to exist for each trait test and statistical tests were done in order to verify this assumption.

Finally candidates were asked to express opinions regarding examination requirements. This expression constituted estimates of face validity.

To estimate construct validity Campbell and Fiske's intercorrelation matrix of multitraits and multimethods was used. To establish the existence of construct validity it was shown that correlations between different methods measuring the same trait should (1) be significantly different from zero, (2) account for a reasonable proportion of test behavior, and (3) should exceed all other correlational values except corresponding reliability estimates. The use of this latter point as

well as an examination of consistent intertrait relationships was described as discriminant validity.

Since Campbell and Fiske's model is a conceptual appraisal, Jöreskog's procedure of a restricted factor analysis was also cited as a suitable technique for determining the existence of constructs (as well as the effect of method variance) in test behavior.

The use of Campbell and Fiske's as well as Jöreskog's models indicated that this study's examinations in general had not elicited behaviors which could be described as factual, comprehension, and problem solving. Two exceptions to this conclusion were observed--namely simulation seemed to elicit some problem solving behavior, and factual-problem solving traits seemed to be the least related traits. Furthermore, candidates in general considered the simulation problems as having face validity for measuring problem solving behavior. These positive results must be cautiously interpreted because (1) simulation tests were the only tasks viewed as relevant, (2) they were also regarded as most difficult, and (3) the technique was certainly more novel than either the essay or multiple choice formats.

The factor analytic solutions reported constrained the general ability factor to be orthogonal to all method and trait factors. Since the α coefficient for all 18 subtests indicated that these tests had high internal consistencies, it is reasonable to assume that this general factor is probably highly correlated with trait factors, and possibly with method factors as well.

This author concludes therefore that the majority of test behavior exibited in over 250 man hours of graduate achievement performances was

not best described by traits labelled as factual, comprehension, and problem solving. With this conclusion what recommendation does this study have for educational measurements and decision making? For that discussion the reader is referred to the final section of this chapter.

Recommendations

Campbell and Fiske (1959, p. 104) indicate that inability to establish reasonable convergent validity estimates is common in studies similar to this investigation. If the validation process is properly regarded as an ongoing process, however, "...validity coefficients obtained at any one stage in the process [should] be interpreted in terms of gains over preceding stages and as indicators of where further effort is needed [Campbell and Fiske, 1959, p. 104]." The following recommendations are made within this framework.

All subtests within all methods seemed to elicit some common factor. This author assumes it was the content of applied statistics and research design. If this assumption is correct, any method used in this study is useful to an evaluator whose objective of measurement is criterion oriented.

Criterion referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement. Accordingly the degree of competence attained by a particular candidate is treatly independently

and references to other performances is needless (Glaser, 1965). Emphasis on the attainment of content rather than on an attainment of operationalized content is becoming more attractive to educators and evaluators. (1) "Development of thinking or problem solving ability can also be considered an objective of schooling...although it is a lesser objective than the learning of subject-matter...[Ausubel, 1963, p. 13]." (2) "...it seems reasonable to conclude that the major goal of education is to develop in the scholars a command of substantive knowledge...[Ebel, 1965, pp. 38-39]." (3) "...schools exist primarily for the purpose of transmitting accumulated knowledge...it is the responsibility of the educator to abstract from this accumulation those elements that are of greatest significance and organize them into teachable units...[Coffman, 1969, p. 4]." (4) "...we should concentrate more on...subject matter or content--to achieve an adequate sampling of the most useful knowledge [Ebel, 1969, p. 74]."

Do the results of this study lead one to concur with these latter quotations? This author believes so. Firstly, there is little doubt that the tests used in this study met or exceeded normal standards of graduate achievement measures. This, coupled with the constraints of the experimental situation, still failed to elicit operationalized behavior which was defined as the objective of measurement. The implication is obvious: most graduate examinations in the area of applied statistics and research design (if not most achievement examinations in general) probably also fail to elicit the three traits

previously defined. Nor is it unreasonable to speculate that many future examinations will also fall short of such an objective. Speculation that validity would have been discovered if Bloom's traits had been used is not very reasonable. If anything the results of this study indicate that at best a dichotomy of traits is only measureable with present day techniques. Even here the two traits were highly correlated.

The suggestion is not being made that mental behavior does not exist as described by Bloom or this study. However, Cronbach (1969) points out that "An item <u>qua</u> item cannot be matched with a single behavioral process. Finding the answer calls for dozens of processes... [p. 43]." In other words the score on a task indicates that the person does or does not possess, in conjunction, <u>all</u> the abilities required to perform it successfully. Nearly forty years ago Thomson (1935) maintained the same point of view: "...the mind is not divided up into 'unitary factors' but is a rich, comparatively undifferentiated complex of innumerable influences [p. 182]." Thus the patterns of two related traits "discovered" in this investigation do not necessarily exist in the manner reported. The discovery may be essentially a definition provided by examinations and statistical models used in this study. Ebel (1969) maintains cognitive complexities have the same characteristics as a biological taxonomy, and consequently cannot be developed or evaluated as abstract entities. It is not unreasonable to assume, however, that the cognitive complexities can be enlarged, or that

elements in the cluster can be strengthened. How can this be accomplished and what bearing will this have for educational measurement? Ebel (1969) sheds some light on these latter topics.

To strengthen cognitive complexities Ebel maintains knowledge (in particular, verbal knowledge) should be taught and evaluated. What does he mean by the term knowledge? It is not simply new information. It is that information which the individual has proven to himself to be the truth. It is not simply a body of concepts. It is a structure which gives facts some degree of coherence and thus makes these facts meaningful and useful. It is not a definitive discipline. It is a class of relevant information.

How is knowledge taught? Give credible answers to students' questions such as "What do you mean?", "How do you know?", or "Why is it so?". How is knowledge evaluated? Compose test items which ask the same questions the students originally asked. How does one compose such items? Continue to use the guidelines of Bloom's hierarchy. If finding an answer for an item calls for a variety of behavioral processes, items written according to Bloom's guidelines will help insure that a variety of mental complexities are tapped. What criteria must items meet in order to be useful? Firstly, they must meet the rigors accumulated research has shown to be necessary. That is to say, questions should be clearly worded, specify a consistent task to all subjects, carefully marked. In essence, the items must be internally consistent. Secondly, and sufficiently, the information tested must, at all costs, be considered by its composer

to be that which he regards as important. In other words, validity is determined by the specialist, not by the performance of the candidate.

The ramifications of this approach are quite simple. (1) It matters little what skill is required to answer a question; what does matter is that the candidate knows the answer. (2) The evaluator is explicity baring his value system. The information he teaches, tests, and upon which he makes educational decisions for selection, placement, or promotion is now open to criticism. It is to be noted, however, this approach demands the specialist to clearly enunciate his educational objectives. As previously stated, the only recourse to progress in education is to expose deficiency. Vague generalities do not allow this exposure, and this author readily admits that even the three traits used in this study can be used as a camouflage for poorly developed instructional and/or evaluative procedures. It has been shown, however, that this study's evaluation techniques did not readily measure these traits and therefore one more deficiency and camouflage has been exposed.

Consequently it is suggested the validity emphasis should change to measuring what Ebel calls knowledge--with little regard to the operationalized use of content. Criticisms that knowledge alone (1) does not guarantee wisdom, (2) will not keep pace with the information explosion, and (3) is usually forgotten, can be refuted to some degree. Firstly, relevant knowledge contributes substantially to wisdom. Secondly, despite the information explosion in specialized areas, little of the

new material affects the quality of most personnel in the field. Thirdly, useful information that is well integrated into a structure of previously accumulated data is not easily forgotten.

In summary, those of us who have taught fully realize there are central cores to each subject, which we regard as fundamentally important to any content area. If one chooses to teach and test in these areas, surely our specialized knowledge and prior experience will allow us to defend out objectives. We will either live with assurance that our procedures were valid, or we will, because of reasonable criticism, change our behaviors. Since the degree of assurance or criticism is highly dependent upon the explicitness of our statement of objectives, the only recourse to having collective agreement on the validity, that is collective agreement on relevancy, is to make our education objectives unambiguously clear so all may inspect and subsequently accept or reject them. Is it not reasonable therefore to de-emphasize our search for the essential traits of mental skill and concentrate upon relevant information which is by definition valid. This author highly recommends such an approach; that is to say, in future investigations researchers would be well advised to concentrate their efforts in the area of content, not construct, validity in order that the utility of achievement examinations may be enhanced.

SELECTED REFERENCES

Adams, G. S. Measurement and evaluation in education, psychology, and guidance. New York: Holt, Rinehard, and Winston, 1967.

All India Council for Secondary Education. Evaluation in secondary schools. New Delhi:

Allard, R. Assessing the consistency of essay markers. Unpublished manuscript, University of Alberta, 1970.

Ammons, M. An empirical study of process and product in curriculum development. Journal of Educational Research, 1964, 57, 451-457.

Ammons, M. Objectives and outcomes. In Encyclopedia of Educational Research (4th Ed.), 1969, 908-914.

Anderson, D. W. Strategies of curriculum development: Works of Virgil Herrick. Merrill, 1965.

Andrew, A. S. Multiple choice and essay tests. Improving College and University Teaching, 1968, 16, 61-66.

APA, AERA NCME. Standards for Educational and Psychological Tests and Manuals. Washington, D.C.: APA, 1966.

Ashbough, E. J. Reducing variability of teachers marks. Journal of Educational Research, 1924, 9, 185-198.

Atkin, J. M. Some evaluation problems in a course content improvement project. Journal of Research in Science Teaching, 1963, 1, 129-132.

Atkin, J. M. Behavioral objectives in curriculum design. A cautionary note. In. R. C. Anderson, et al., Current Research on Instruction. Englewood Cliffs: Prentice-Hall, 1969, 60-65.

Ayers, J. D. Summary Description of Grade Nine Social Studies Objectives and Test Items. Edmonton: Department of Education Examinations, 1966.

Ausubel, D. P. The psychology of meaningful verbal learning. New York: Grune and Stratton, 1963.

Baldwin, T. S. Evaluation of learning in industrial education. In B. S. Bloom, J. T. Hastings, & G. F. Madaus (Eds.), Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Bechtoldt, H. P. Construct validity: a critique. American Psychologist, 1959, 14, 619-629.

Bialek, H. M. A measure of teacher's perceptions of Bloom's educational objectives. Paper read at the annual meeting of the American Educational Research Association, New York, February, 1967. Cited by H. J. Sullivan, Objectives, evaluation, and improved achievement. AERA Monograph Series on Curriculum Evaluation (3). Washington, D.C. Rand McNally, 1969.

Bloom, B.S., Engelhart, H.D., Furst, E.J., Hill, W.H. & Krathwohl, D.R. Taxonomy of educational objectives, Handbook I: Cognitive domain. New York: David McKay, 1956.

Bloom, B.S., Hastings, J.T. & Madaus, G.F. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Bruner, J. S. The process of education. Harvard University Press, 1960.

Blumer, G. Desirability of changing the type of written examinations. Journal of the American Medical Association, 1919, 72, 1131-1133.

Boruch, R. F., Larkin, J. D., Wolins, L. & MacKinney, A. C. Alternative methods of analysis: Multitrait-multimethod data. Educational and Psychological Measurement, 1970, 30 (4), 833-53.

Boruch, R. F. & Wolins, L. A procedure for estimation of trait, method and error variance attributable to a measure. Education and Psychological Measurement, 1970, 30 (3), 547-74.

Bridge, E. M. External examinations in medical science. Journal of Medical Education, 1956, 31, 174-180.

Brogden, H. E. Variation in test validity with variation inthe distribution of item difficulties, number of items, and degree of their intercorrelation. Psychometrika, 1946, 11, 197-214.

Brooks, C. Personal communication cited by J. P. Hubbard & W. V. Clemans, Multiple-choice examinations in medicine. Philadelphia:

Bull, G. M. Examinations. Journal of Medical Education, 1959, 34, 1154-1158.

Campbell, D. J. Recommendations for APA test standards regarding construct, trait and discriminant validity. American Psychologist. 1960, 15, 546-553.

Campbell, D. T. & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56 (2), 81-105.

Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Cason, H. The essay examination and the new type of test. School and Society, 1931, 34, 413-418.

Cast, B. M. D. The efficiency of different methods of marking English compositions. British Journal of Educational Psychology, 1939, 9, 257-269.

Cast, B. M. D. The efficiency of different methods of marking English compositions. British Journal of Educational Psychology 1940, 10, 49-60.

Centra, J. A. Validation by the multigroup-multigroup-multiscale matrix: An adaptation of Campbell and Fiske's convergent and discriminant validation procedure. Educational and Psychological Measurement, 1971, 31, 675-683.

Coffman, W. E. Concepts of achievement and proficiency. Proceedings of the 1969 invitational conference on testing problems. Princeton: Educational Testing Service, 1969.

Coffman, W. E. & Kurfman, D. A comparison of two methods of reading essay examinations. Journal of Educational Research, 1968, 5, 99-107.

Cowles, J. T. & Hubbard, J. P. A comparative study of essay and objective examinations for medical students. Journal of Medical Education, 1952, 27, 14-17.

Cowles, J. T. Current trends in examination proceedings. Journal of American Medical Association, 1954, 155, 1383-1387.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-234.

Cronbach, L. J. Essentials of psychological testing 2nd Ed. New York: Harper & Row, 1960.

Cronbach, L. J. Validation of education measures. Proceedings of 1969 invitational conference on testing problems. Princeton: Educational Testing Service, 1969.

Cronbach, L. J. & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin. 1955, 52, 281-302.

Darsie, M. L. The reliability of judgements based on the willing compositic scale. Journal of Educational Research, 1922, 5, 89-90.

Davis, O. L. & Tinsley, D. Cognitive objectives revealed by classroom questions asked by social studies student teachers. Paper read at the annual meeting of the American Educational Research Association, New York, February, 1967. Cited by W. J. Popham, Objectives and instruction. AERA Monograph Series on Curriculumn Evaluation (3) Washington D.C.: Rand McNally, 1963.

Dewey, J. The school and society. University of Chicago Press, 1915.

Dressel, P. L. Evaluation in general education. Dubuque: Brown, 1954.

Dressel, P. L. Evaluation in higher education. Boston: Houghton Mifflin, 1961.

Dyer, H. S. Educational measurement--its nature and its problems. In Harry D. Berg (Ed.), Evaluation in social studies. 35th Yearbook of the National Council for the Social Studies, 1965.

Ebel, R. L. Content standard test scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Ebel, R. L. Measuring educational achievement. Englewood Cliffs: Prentice-Hall, 1965.

Ebel, R. L. Knowledge vs. ability in achievement testing. Proceedings of 1969 invitational conference on testing problems. Princeton: Educational Testing Service, 1969.

Edwards, P. D. M. Symposium: The use of essays in selection at 11.I. Essay marking experiments: Shorter and longer essays. British Journal of Educational Psychology, 1956, 26, 128-136.

Eells, W. t. Reliability of repeated grading of essay type of examinations. Journal of Educational Psychology, 1930, 21, 48-52.

Eisner, E. W. "Educational objectives: Help or hindrance?" The School Review, 1957, 75 (3), 250-276.

Eisner, E. W. Instructional and expressive objectives: Their formulation and use in curriculum. AERA Monograph Series on Curriculum Evaluation (3). Washington, D. C. Rand McNally, 1969, 1-18.

Fastier, F. N. Consistency of performance at examinations. Journal of Medical Education, 1959, 34, 761-72.

French, W. Behavioral goals of general education in high school. New York: Russell, 1957.

Friedenberg, E. Z. Social consequences of educational measurement. Proceedings of the 1969 Invitational Conference in Testing Problems. Princeton: Educational Testing Service, 1969.

Furst, E. J. Constructing evaluation instruments. New York: Longmans, Green, and Co., 1958.

Gagné, R. M. & Bolles, R. C. A review of factors in learning efficiency. Automatic Teaching: The State of the Art. New York: Weley, 1959.

Gagné, R. M. The acquisition of knowledge. Psychological Review, 1962, 69, 335-365.

Gagne, R. M. Learning and proficiency in mathematics. Mathematics Teacher, 1963, 56, 623.

Gagne, R. M. The conditions of learning. New York: Holt, Rinehart and Winston, 1965.

Gagne, R. M. The analysis of instructional objectives for the design of instruction. In R. Glaser Teaching machines and programmed learning: Data and directions. Washington, D. C.: National Education Association of the United States, 1965, 21-65.

Gagne, R. M. Curriculum research and the promotion of learning. AERA Monograph 1, Perspectives of Curriculum Evaluation. Chicago: Rand McNally, 1967.

Goldstein, A. An inquiry into the value of rank grades in the medical course. Journal of Medical Education, 1958, 33, 193-200.

Grant, D. L. & Caplan, N. Studies in the reliability of the short-answer essay examination. Journal of Educational Research, 1957, 51, 109-16.

Green, E. J. The learning process and programmed instruction. New York: Holt, Rinehart and Winston, 1962.

Gronlund, N. E. Stating behavioral objectives for classroom instruction. New York: MacMillan, 1970.

Guilford, J. P. The nature of human intelligence. New York: McGraw-Hill, 1967.

Gulliksen, H. Theory of mental tests. New York: John Wiley, 1950.

Harasym, P. Effects of three cognitive levels of questions on achievement. Unpublished thesis, University of Alberta, 1970.

Hartog, P. J. An examination of examinations. London: MacMillan, 1935.

Hazlett, C. B. The storage and retrieval of multiple choice items on computer. Unpublished thesis, University of Alberta, 1969.

Hazlett, C. B. Research and Information Report: MEDSIRCH. Edmonton, University of Alberta, DERS-3-70, 1970.

Hazlett, C. B., Maguire, T. O. & Wilson, D. R. Inter-judge reliability on essay examinations. Paper read at the 27th Annual Conference of Canadian Medical Colleges, Toronto, 1969.

Hilgard, E. R. Introduction to Psychology 2nd Ed. New York: Harcourt & Brace, 1957.

Hoffman, B. The tyranny of testing. New York: Collier McMillian, 1962.

Holland, J. G. Teaching machines: fundamental principles. In D. S. Schechter (Ed.), Programmed instruction and the hospital. Chicago: Hospital research and educational trust, 1967, 50-61.

Hoyt, C. J. Test reliability obtained by analysis of variance. Psychometrika, 1941, 6, 153-160.

Hubbard, J. P. & Clemans, W. V. Multiple choice examinations in Medicine. New York: Len and Febiger, 1961.

Hubbard, J. P., Lenet, E. S., Schumaker, C. F., & Schnakel Jr., T. G. An objective evaluation of clinical competence. New England Journal of Medicine, 1965, 272, 1321.

Hubbard, J. P. and Clemans, W. V. A comparative evaluation of medical schools. Journal of Medical Education, 1960, 35, 134-141.

Huddleston, E. M. Measurement of writing ability at the college entrance level: Objective vs. subjective testing techniques. Journal of Experimental Education, 1954, 22, (3).

Hulton, C. E. The personal element in teachers marks. Journal of Educational Research, 1925, 12, 49-55.

Jenkins, J. J. & Lykken, D. T. Individual differences. Annual Review of Psychology, 1957, 8, 79-112.

Jessor, R. & Hammond, K. R. Construct validity and the Taylor Anxiety Scale. Psychological Bulletin, 1957, 54, 161-170.

Joreskog, K. G. Some contributions to maximum likelihood factor analysis. Psychometrika, 1967, 32, 443-82.

Joreskog, K. G. A general method for analyses of covariance structures. Biometrika, 1970, 57 (2), 239-51.

Joreskog, K. G. Statistical analysis of sets of congeneric tests. Psychometrika, 1971, 36 (2), 109-133.

Joreskog, K. G. & Gruvaeus, G. RMLFA, a computer program for restricted maximum likelihood factor analysis. Research Bulletin. Princeton: Educational Testing Service, 1967, 21.

Joreskog, K. C., Gruvaeus, G. T. & van Thello, M. A general computer program for analysis of covariance structures. Research Bulletin. Princeton: Educational Testing Service, 1970, 15.

Karsner, H. T. Personnal communication cited by J. P. Hubbard & W. V. Clemans, Multiple-choice examinations in medicine. Philadelphia: Leo & Febiger, 1961, 93-4.

Kearney, N. C. Elementary school objectives. New York: Russell, 1953.

Krathwohl, D. R., Bloom, B. S., and Masia, B. B. Taxonomy of educational objectives: Handbook II, the affective domain. New York: David McKay, 1964.

Kropp, R. P., & Stoker, H. W. The construction and validation of tests of the cognitive processes as described in the "Taxonomy of educational objectives". Cooperative Research Project No. 2117. Tallahasses: Florida State University, Institute of Human Learning, 1966. Cited by H. J. Sullivan, Objectives evaluation, and improved achievement. AERA Monograph Series on Curriculum Evaluation (3). Washington, D. C.: Rand McNally, 1963.

Kuder, G. F., & Richardson, M. W. The theory of estimation of test reliability. Psychometrika 1937, 2, 151-160.

Lawrence, G. D. Analysis of teacher-made tests in the social sciences according to the "Taxonomy of Educational Objectives". Masters thesis, Claremont College, 1963. Cited by H. J. Sullivan, Objectives, evaluation, and improved achievement. AERA Monograph Series on Curriculum Evaluation (3). Washington, D. C.: Rand McNally, 1963.

Lennox, B. A comparative trail of objective papers and essay papers in pathology and bacteriology class examinations, The Lancet, August 31, 1957, 396-402.

Levine, H. G. & McGuire, C. H. The validity and reliability or oral examinations in assessing cognitive skills in medicine. Journal of Educational Measurement, 1970, 1 (2), 63-74.

Levine, H. G., McGuire, C. H., & Nattress, L. W. The validity of multiple choice achievement tests as measures of competence in medicine. American Educational Research Journal, 1970, 7, (1), 69-82.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. Psychological Monograph, 1947, 61 (4).

Loevinger, J. Objective tests as instruments of psychological theory. Psychological Reports, Monograph Supplement 9, 1957, 636-694.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading: Addison-Wesley, 1968.

Lumsdaine, A. A. Instructional research: some aspects of its status, defects and needs. In H. I. Klausmeier and G. T. O'Nearn (Eds.), Research and development toward the improvement of education. Madison: Dembar Educational Research Services, 1968. 95-101.

MacDonald, J. B. Myths about instruction. Ed Leadership, 1965, 22, 571-576.

Mager, R. F. Preparing objectives for programmed instruction. San Francisco: Fearon, 1962.

Maguire, T. O. & Hazlett, C. B. Reliability for the researcher. Alberta Journal of Educational Research, 1969, 15 (2), 117-126.

Manning, W. H. The functions and uses of educational measurement. Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton: Educational Testing Service, 1969.

Marschak, J. Probability in the social sciences. In P. F. Lazarsfeld (Ed.), Mathematical thinking in the social sciences. Glencoe: Free Press, 1954, 116-215.

Marshall, M. S. Objections to the objective objective. The Educational Forum, 1956, 279-85.

McCallough, A. & Flanagan, B. The validity of machine-scorable cooperative English test. Journal of Experimental Psychology, 1939, 7, 229-234.

McDonald, G. W. Programming for patient and professional education in diabetes. In D. S. Scheckter (Ed.), Programmed instruction and the hospital. Chicago: Hospital research and educational trust, 1967, 105-113.

McGuire, C. H. & Babbott, D. Similation technique in the measurement of problem-solving skills. Journal of Educational Measurement, 1967, 4 (1), 1-10.

Michael, J. J. "Structure of intellect theory and the validity of achievement examination". Educational and Psychological Measurement, 1968, 28, 1141-9.

Miller, G. E., McGuire, C. H., Larsen C. B. The orthopedic training study. Bulletin of the American Academy of Orthopaedic Surgeons, 1965, 13, 8-11.

Moore, R. A. Methods of examining students in medicine. Journal of Medical Education, 1954, 29, 1.

Moore, W. J. & Kennedy, L. D. Evaluation of learning in the language arts. In B. S. Bloom, J. T. Hastings, G. F. Madaus (Eds.), Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

National Education Association. The central purpose of American education. NEA, 1961.

Nyberg, V. R. The reliability of essay grading. Paper read at the 6th Annual Canadian Conference on Educational Research, Edmonton, 1966.

Orlandi, L. R. Evalution of learning in secondary schools social studies. In B. S. Bloom, J. T. Hastings, G. F. Madaus (Eds.), Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Paterson, D. G. Do new and old examinations measure different functions? School and Society, 1926, 34, 246-8.

Pidgeon, D. A & Yates, A. Symposium: Use of essays in selection at 11. IV. Experimental inquiries into the use of essay type english papers. British Journal of Educational Psychology ,1957, 27, 37-47.

Popham, W. J. Objectives and instruction. AERA Monograph Series on Curriculum Evaluation (3). Washington, D. C.: Rand McNally, 1969(a).

Popham, W. J. Probing the validity of arguments against behavioral goals. In R. C. Anderson, G. W. Faust, M. C. Roderick, D. J. Cunningham, & T. Andre (Eds.), Current Research on Instruction. Englewood Cliffs: Prentice-Hall, 1969(b).

Purves, A. C. Evalution of learning in literature. In B.S. Bloom, J. T. Hastings, G. F. Madaus (Eds.), Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971

Romey, W. D. Inquiry techniques for teaching science. Englewood Cliffs: Prentice-Hall, 1968.

Scannell, D. P. & Stellwagon, W. R. Teaching and testing for degrees of understanding. California Journal of Instructional Improvement, 1960, 3, 1.

Scheffler, I. Philosophical models of teaching. Harvard Educational Review, 1965, 35, 131-143.

Scott, J. C. & Burke, N. B. An automatic examination machine for medical students. Journal of Medical Education, 1957, 32, 427-432.

Skinner, B. F. Teaching machines. Science, 1958, 128, 969-977.

Stalnaker, J. M. Essay examinations reliably read. School and Society, 1937, 46, 671-2.

Stalnaker, J. M. The essay examination. In E. F. Lindquist (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1951.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), Educational Measurement. Washington D.C.: American Council on Education, 1971.

Steele, J. M. Dimensions of the class activities questionnaire. Urbana: University of Illinois, 1969.

Steele, J. M. Assessing intent and practice in instruction. Paper read at the American Educational Research Association, March, 1970.

Stolurow, L. M. Teaching by machine. Washington, D. C.: U. S. Department of Health, Education, and Welfare, 1961.

Stolurow, L. A. Socrates of a computer-based instructional system in theory and research. In H. I. Klausmeier and G. T. O'Hearn (Eds.), Research and development toward the improvement of education. Madison: Dembar Educational Research Services, 1968, 102-117.

Sullivan, H. J. Objectives, evaluation and improved learner achievement. AERA Monograph Series on Curriculum Evaluation (3), 1969, 65-97.

Sullivan, H. J. Objectives evaluation, and improved learner achievement. AERA Monograph Series on Curriculum Evaluation (3). Washington, D.C.: Rand McNally, 1969.

Taba, H. Curriculum development, theory and practice. New York: Harcourt, Brace, and World, 1962.

Taber, J. I., Glaser, R., & Schaefer, H. H. Learning and programmed instruction: Reading: Addison-Wesley, 1965.

Thomson, G. H. The factorial analysis of human abilities. Human Factor, 1935, 9, 180-185.

Thorndike, R. L. Reliability. In E. F. Lindquist (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1951, 560-620.

Tryon, R. C. Reliability and behavior domain validity: Reformation and historical critique. Psychological Bulletin, 1957, 54, (3), 229-249.

Tyler, L. L. A case history: Formulation of objectives from a psychoanalytic framework. AERA Monograph Series on Curriculum Evaluation (3). Washington, D.C.: Rand McNally, 1969.

Tyler, R. W. Basic principles of curriculum and instruction. University of Chicago, 1950(a).

Tyler, R. W. The functions of measurement in improving instruction. In E. F. Lindquist (Ed.), Educational measurement. Washginton, D. C.: American Council on Education, 1950(b).

Tyler, R. W. The fact-finding study of the testing program of the United States Armed Forces Institute, 1952-1954. Report to the USAFI, University of Chicago, 1954.

University Grants Commission. Evaluation in higher education. New Delhi: UGC, 1961.

Valette, R. M. Evaluation of learning in a second language. In B. S. Bloom, J. T. Hastings, G. F. Madaus (Eds.), Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Vernon, P. E. & Millican, G. D. A further study of the reliability of english essays. British Journal of Statisitcal Psychology, 1954, 7, 65-74.

Weidemann, C. C. Further studies of the essay test. Journal of Higher Education, 1941, 8, 437-9.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

Winetrout, K. The national teacher examinations. Journal of Higher Education, 1941, 12, 479-84.

Wiseman, S. Symposium: The use of essays in selection at 11. III. Reliability and validity. British Journal of Educational Psychology 1956, 26, 172-9.

Wiseman, S. & Wrigley, J. Essay reliability: The effect of choice of essay title. Educational and Psychological Measurement, 1958, 18, 129-38.

Wolf, R. Review of "The construction and validation of tests of the cognitive processes as described in the 'taxonomy of educational objectives'." Educational and Psychological Measurement, 1967, 27, 542-548.

APPENDIX A

Multiple Choice Tests

A.1: Multiple Choice 1

DIRECTIONS: For each question, circle the BEST answer.

ID_____

C1 1. If every score in a distributuon of scores is measured as a deviation from some arbitrary point, the sum of the squared deviations will be a minimum when the point selected is:

A. the mean
B. the median
C. the lowest score
D. zero
E. the standard deviation

F1 2. Deviation values are usually defined as

A. $X_i - S$

B. $\overline{X}_i - X_i$

C. $X_i - \overline{X}$

D. $X_i = \overline{X}/S$

E. $X_i - \overline{X}/S^2$

F2 3. Arriving at a conclusion or decision while explicitly recognizing the probability that a wrong conclusion or decision will be made is known as

A. statistical inference
B. classical sampling theory
C. scientific methodology
D. inductive reasoning
E. descriptive statistics

F3 4. When expected frequencies are obtained from the marginal frequencies of a four-fold contingency table, the chi square obtained can be used to test a hypothesis of

A. no relationship
B. equal means
C. equal variances
D. equal population frequencies
E. both A and D

C2 5. The slope of the regression line of A on B is the same as the slope of the regression line of B on A when the correlation coefficient r_{AB} has a value of

A. -1.00
B. 0
C. .50
D. 1.00
E. both A and D

P1 6. Consider research techniques in general rather than a specific problem. Then we can say that nonlinear regression differs from linear regression in that in the former

A. cumulative frequencies are used
B. deviations from the regression line are minimized
C. deviations from the line of means are minimized
D. many different functional relationships between variables can be used
E. both C and D

P2 7. A teacher gives a test of understanding in social science and an achievement test in reading. The scores on the two tests correlate with an r = .90. Which of the following conclusions can best be drawn from the above information?

A. 90 times out of 100 students making a high score in reading will make a high score on a social science test
B. people who understand social science do so because they are good readers
C. there seems to be a relatively high relationship between reading and social science test scores
D. 10 per cent of the students who make a high score on the reading test will score high on the social science test
E. studying social science improves reading test scores

C3 8. Inference is distinguished from pure description

A. by the notion of probability
B. in the method of classification
C. in the use of analogy
D. in the nature of the materials of observation
E. in the neglect of negative cases

C4 9. Which of the following statements is <u>always</u> characteristic of the median?

A. it is a point midway between X_{25} and X_{75}
B. the sum of the deviations of the raw scores from the median is zero
C. a vertical line drawn upward from the median intersects the highest portion of the distribution
D. the distribution could be "balanced" at the median
E. a vertical line drawn upward from the median divides the histogram into two equal areas.

F4 10. When experimental results are due in part to enthusiam for a new treatment, the situation is referred to as the

A. Hawthorne effect
B. Novelty Effect
C. Experimenter effect
D. both A and B
E. both A and C

F6 16. The standard error of a test score is a function of both the reliability of the test and

A. the mean score on the test
B. the standard deviation of scores on the test
C. the normal distribution of standard scores
D. positive skew
E. both B and C

C7 17. In the application of classical statistical theory of the construction of confidence intervals for the population mean, it is assumed that n, the sample size, is "sufficiently large." In this context "sufficiently large" means

A. large enough that $\bar{X}$. has a t distribution
B. greater than 30
C. large enough that the sampling distribution of $\bar{X}$. can be considered normal
D. large enough to make the parent population approximately normal
E. infinitely large

C8 18. If you know the mean and the standard deviation of a population, what additional information do you need in order to describe (approximately) the distribution of random sample means?

A. none
B. the level of confidence
C. the shape of the population distribution
D. the sample size
E. the sample standard deviation

C9 19. In general, grouping data into class intervals does not result in a loss of information for

A. the total frequency
B. the standard deviation
C. median
D. the mode
E. the mean

P5 20. Which of the following topics would lend itself most definitely to experimentation?

A. the interests of children in various kinds of books.
B. the number and duration of eye fixations under different conditions of lighting and type size
C. an analysis of the reading habits of boys and girls of different intellectual levels
D. the reading proficiency of children wearing glasses in contrast to those not requiring glasses
E. the relationship of motivation to academic achievement

F5 11. Chi-square is used with frequency data. Consequently, it is used to compute an index of correlation called the contingency coefficient at what level of measurement?

A. nominal
B. ratio
C. ordinal
D. interval
E. both A and C

C5 12. The numerator of the F ratio for analysis of variance is based on

A. the variation within samples
B. the variation of sample means
C. the total variation of individual cases
D. the variation from characteristic to characteristic
E. both A and C

C6 13. The scores on most standardized achievement tests are

A. nominal scale scores
B. better than ordinal scale scores but not quite interval scale scores
C. interval scale scores
D. better than interval scale scores but not quite ratio scale scores
E. better than nominal scale scores but not quite ordinal scale scores

F3 14. Four graduate students are discussing why there is a requirement in Education for courses in statistical methods. Alice says that it is so that one can describe persons, groups, and events more effectively. Betty says that it is so educators can defend their decisions more effectively. Clara decides that it is so that future educators will have better diciplined minds, free from value judgments. Doris feels that it is really part of the selection process so that just certain people will get advanced degrees in Education. Which student(s) most nearly reflects the point of view of most educational psychologists?

A. Alice
B. Betty
C. Clara
D. Doris
E. all of the above

P4 15. It is known that the population correlation coefficient on a continuous scale is 1.0. A sample of 10 is tested, and the sample correlation coefficient is found to be .95. What can be concluded from this information?

A. the sample was DEFINITELY drawn from the population
B. the sample DEFINITELY was NOT drawn from the population
C. the sample was PROBABLY drawn from the population
D. the sample PROBABLY was NOT drawn from the population
E. to determine if the sample was or was not drawn from the population, MORE INFORMATION is needed.

P6 21. The basic limitation of experimentation as the primary source of scientific data in education is that

A. experimentation is slow and time-consuming
B. not all educational problems are amenable to experimentation
C. many teachers do not know how to conduct scientific experimentation
D. experimentation in education cannot guarantee the control of extraneous variables required to provide valid results
E. ultimate purposes must be defined arbitrarily and not determined by scientific investigation

F7 22. A random sample

A. provides the basis for estimating population parameters
B. gives every measure in the population an equal chance of being selected
C. permits application of the laws of chance
D. A and B above
E. all of A, B, and C

C10 23. The mean and standard deviation are the chief measures of central tendancy and variability used because

A. they are the easiest to calculate from ungrouped data
B. the magnitude of all of the scores is used in their calculation
C. they are least susceptible to misinterpretation
D. most distributions are approximately normally distributed
E. C and D above

C11 24. If a correlation coefficient is judged to be significant

A. it will be useful for prediction
B. it represents a perfect relationship between the variables
C. it will tend to be larger if calculated from a larger number of cases
D. it may have a low numerical value
E. it has a positive relationship

F8 25. When the experimenter generalizes from the results of a particular experiment to the class of all similar experiments, this extension from the particular to the general is called

A. reasoning from population to sample
B. deductive inference
C. inductive inference
D. going from known to the unknown
E. reasoning from hypothesis to observations

P9 26. Discrete attributes are those that can be measured in

A. whole units
B. fractional units
C. intervals
D. percentiles
E. B and C above

P7 27. If you wished to represent the typical salary of education personnel (superintendents, principals, and teachers) in Alberta, you would AVOID using the

A. mean
B. median
C. mode
D. most commonly accuring salary
E. both C and D

P8 28. Common sense inquiry differs chiefly from scientific inquiry in that

A. it does not use facts and ideas which are the product of earlier inquiries
B. the attitudes and habits of earlier experiences operate only in a casual way
C. scientific inquiry takes care to insure the relevancy of the facts and conceptions to be employed in inquiry
D. knowledge attained in common sense inquiry is mainly the problem of use and enjoyment
E. scientific knowledge is not attained for its own sake

C12 29. It is desired to test the hypothesis that the mean of a normal distribution is equal to 50 against the alternative that the mean is not equal to 50. Which of the following plans will incur the least risk of rejecting the null hypothesis if it is true

A. sample size 500 level of significance .05
B. sample size 50 level of significance .01
C. sample size 100 level of significance .02
D. Sample size 1000 level of significance .10
E. either A or D

C13 30. If the sample size in a test of a statistical hypothesis is kept constant, and the probability of making a type I error (α) is increased from .05 to .25, the probability of making a type II error (β) is

A. always increased
B. always decreased
C. unaffected
D. usually increased
E. usually decreased

C14 31. The "regression effect" will be greatest for which of the following correlations

A. -0.5
B. -0.1
C. 0.05
D. +.5
E. 1.00

F10 32. The variance of the set of scores (4,4,4,4,4,4,) is

A. 0
B. 1
C. 4
D. 16
E. 96

C15 33. The confidence interval for a sample mean does <u>not</u> depend on which of the following

A. the standard deviation of sample scores
B. the size of the sample
C. the value of the population mean
D. the confidence level decided upon
E. both B and D

P9 34. Hypotheses in research serve the following purpose:

A. colligating the facts observed about different phenomena into some simple form
B. determining the validity of alternative hypotheses
C. anticipation of the statistical techniques required for dealing with the problem
D. forcing recognition of underlying assumptions
E. limiting the field of investigation

C16 35. Two scores are added to a normal distribution, one at the 30th percentile and the other at the 95th perentile. We can expect that the mean of the group will _____ and the median score will _____.

A. increase; also increase
B. decrease; also decrease
C. remain constant; also remain constant
D. remain constant; increase
E. increase; remain constant

F11 36. The probability of rejecting a null hypothesis which is in fact true is given by

A. the significance level
B. the β error
C. dependent on the alternate hypothesis
D. the F distribution
E. level of confidence

C17 37. Why is it sensible to construct a "scattergram" in the calculation of a product-moment correlation coefficient even if the data are to remain ungrouped?

A. to facilitate computation
B. to predict one variable from another
C. to display relationships between the means and the standard deviations
D. to ascertain if the correlation will be high enough to justify calculation
E. to ascertain if the relationship is approximately linear

F12 38. A continuum characteristic

A. is based on differences in amount
B. is based on differences in kind
C. has meaningful units that are equal in size at different parts of the scale
D. has a zero point and positive scores but no negative scores
E. all of A, B, and C

F13 39. A distribution of sample statistics, the same statistic taken from each of many samples drawn at random from the population, is known as

A. the distribution of a sample
B. a distribution of estimators
C. the population distribution
D. a sampling distribution
E. none of the above

C18 40. In a <u>very</u> negatively skewed frequency distribution, one is likely to find

A. few cases at the mean
B. more than half the cases below the mean
C. bimodality
D. many negative scores
E. little kurtosis

F14 41. The sample mean is a

A. statistic
B. characteristic
C. frequency
D. parameter
E. both A and C

P10 42. A research assistant does a pooled variance t-test but should have done a correlated t-test. The value of "t" he obtained will be

A. appropriate
B. smaller than it should have been
C. larger than it should have been
D. same, or smaller, than it should have been
E. inappropriate, but nothing can be said as to its relative size

P11 43. For 1043 industrial arts majors the Pearson r between aptitude and performance scores is found to be about =.60. Which of the following statements is NOT appropriate?

A. in the population that this sample represents the correlation between aptitude and performance is probably not zero
B. in industrial arts, high performance is dependent on the kind of aptitude measured by the test that was used.
C. about 35% of the variability in performance scores can be attributed to variability in aptitude scores
D. industrial arts majors who have high aptitude scores tend to be above the mean in performance
E. both A and B

F15 44. Before a statistical investigation begins, one should

A. calculate mean and σ of the population
B. define the population
C. define the units of measurement
D. send out a questionnaire
E. randomly select his sample

P12 45. In his class Johnny's score on test A was at the 95th percentile, his score on test B was also at the 95th percentile. If each student in Johnny's class had his scores from test A and test B added together, at what percentile would Johnny's combined score probably be amongst all the combined scores, IF TEST A and B WERE CORRELATED 1.0?

A. less than 95P
B. equal to 95P
C. greater than 95P
D. such a condition could not exist
E. need more information before any reasonable guess could be made

A.2: Multiple Choice 2

DIRECTIONS: For each question, circle the BEST answer. ID_____

F1 1. "Statistically significant" is often used to refer to differences between

A. values of population parameters
B. values of sample statistics
C. probabilities of type I and type II errors
D. probabilities of $1-\beta$ and $1-\alpha$
E. all of the above

F2 2. The essential step in the experimental procedure which serves as a safeguard for valid interpretation of experimental results in spite of the impossibility of securing uniformity of conditions not under experimental control is the application of the principle of

A. replication
B. control of variation
C. randomization
D. a self-contained experiment
E. experimental inference

F3 3. Descriptive statistics

A. are verbal interpretations of tables
B. include measures of central tendency
C. include measures of dispersion of scores
D. summarize data
E. both B and C

F4 4. A population

A. refers to the number of people in a sample
B. is a collection of measurements
C. is specified after analysis of samples
D. is a universe of all possible measures of a certain kind
E. is both B and D

F5 5. The halo effect is mainly a result of the fact that raters are influenced by

A. their overall appraisal of the individual being rated
B. the ratings made by others on the same individual
C. the tendency to rate most people above average
D. the tendency to give a rating of "average" for any trait on which they are not sure
E. the tendency to show favoritism toward certain individuals

F6 6. The standard error of the mean

A. is avoidable by careful sampling
B. is a measure of the dispersion of a large number of means of samples taken from the same population
C. becomes smaller when the number of each sample is decreased
D. has no use in inferential statistics
E. is both A and B

C1 7. When scores of two or more groups are pooled

A. the total N is the sum of the Separate N's
B. the mean of all pooled scores is the average of all the groups means
C. the variance of pooled scores is the sum of the separate group variances
D. all of the above
E. both A and B

C2 8. In a study involving 5,000 cases one characteristic was reported to have a mean of 98, a median of 90, and a mode of 101. If no mistakes had been made in computation, it would be reasonable to suppose that

A. the distribution was symetrical
B. the shape of the distribution was quite irregular
C. the sample was very homogeneous
D. percentile ranks should have been used
E. the distribution was positively skewed

P1 9. The chief defect of a systematic sample is that

A. it is more difficult to select than a simple random sample
B. it does not give an unbiased estimate of the mean
C. it does not give every unit of the population an equal probability of inclusion
D. its results do not make possible any formula of general validity for the sampling error of the estimate.
E. it is too difficult to select where the drawing must be done in the field

F7 10. Standard scores between $z = \pm 1.96$ include ABOUT what percent of the area under a unit normal distribution?

A. 1%
B. 2.5%
C. 5%
D. 95%
E. 99%

F8 11. Chi square is a distribution most frequently used to test hypotheses about _____data

A. random
B. graduated
C. enumeration
D. measurement
E. both C and D

P2 12. Suppose you have two intact classes available, both taught by the same teacher, and you wish to test the efficacy of a particular visual aid in instruction. For each unit through the semester the aid is assigned randomly to one class or the other. At the end of the semester, the units taught by visual aid are significantly superior to those taught without. You can generalize the results to presence of the aid in general under which of the following assumptions?

A. no interaction exists between treatments and any of the specific variables such as teacher, class, etc.
B. this teacher is an average teacher and the class an average class
C. the classes, though intact are essentially equivalent
D. both A and B
E. both B and C

F9 13. In a prediction study where A is predicted from B, the variable A is

A. a stimulus variable
B. a dependent variable
C. an independent variable
D. an intervening variable
E. both B and D

P3 14. In an experiment, the investigator attempts to have a control group which is

A. identical in all respects to the experimental group
B. identical in all respects to the experimental group, except for the factor under study
C. different in all respects from the experimental group
D. opposite in all respects to the experimental group except for the factor under study
E. identical in all relevant respects to the experimental group except for the factor under investigation

C3 15. A researcher using an $\alpha = .05$, carried out 126 independent t-tests. If Ho WERE TRUE, approximately how many significant t's would you estimate this researcher would find:

A. none
B. 1
C. 5
D. 6
E. no estimate can be made with any degree of probability

C4 16. In which of the following distributions is the greatest number of cases below the mean?

A. negatively skewed
B. positively skewed
C. bimodal
D. normal
E. rectangular

P4 17. A worker could have used a t-test on his data but decided to do a median test instead. The switch

A. will have no effect on his conclusion
B. might cause a Type I error
C. might cause a Type II error
D. violated the basic assumptions of statistical inference
E. does all of B, C, and D

P5 18. It is known that the population correlation coefficient on a continuous scale is -1.0. A sample of 4 is drawn from this population, and the sample correlation coefficient is calculated. The value of this sample correlation coefficient

A. would PROBABLY be negative
B. COULD be positive
C. COULD be zero
D. all of the above
E. none of the above

C5 19. The most useful method of sampling is

A. random sampling with replacement
B. random sampling without replacement
C. statified sampling
D. systematic sampling
E. either C or D

F10 20. A normal curve has no

A. mode
B. continuum characteristics
C. skewness
D. specific formula
E. kurtosis

C6 21. In a descriptive study, one of the primary considerations is

A. formulating hypotheses that are capable of being tested
B. selecting a sample that has the same characteristics as the whole population
C. deciding upon what data to gather
D. keeping bias to a minimum
E. both C and D

F11 22. The expected value of the random variable X is the same as the

A. sample mean of X
B. population mean of X
C. median value of X
D. most probable value of X
E. none of the above

P6 23. If a given raw score is determined to have a percentile rank of 40 and a z - score of +.10, it follows that

A. the distribution is negatively skewed
B. the distribution is approximately normal
C. the distribution is positively skewed
D. distribution is irregular in some way
E. those scores were incorrectly computed and assigned

P7 24. The 95 percent confidence interval on the difference between the means $\bar{X}._1$ and $\bar{X}._2$ does not include 0. If the null hypothesis $\mu_1 - \mu_2 = 0$ with its alternative $\mu_1 - \mu_2 \neq 0$ were now tested, which of the following would be true?

A. $\bar{X}._1$ and $\bar{X}._2$ would be significantly different at the .025 level

B. $\bar{X}._1$ and $\bar{X}._2$ would not be significantly different at the .05 level

C. $\bar{X}._1$ and $\bar{X}._2$ would be significantly different at the .05 level

D. $\bar{X}.$ and $\bar{X}._2$ would be significantly different at the .10 level

E. both C and D

P8 25. Students enrolled in physics in a particular high school were randomly divided into two groups of 20 pupils each. One of the groups was taught by an inductive method of instruction, and the other by a deductive method. This differentiated instruction was carried on for a period of six weeks during which the same teacher instructed both groups in a unit on electricity. A unit test of satisfactory reliability and validity was administered to all of the students before and at the conclusion of the period of differentiated instruction.

Consider each of the following hypotheses and indicate which is (or are) NOT TESTABLE in the above investigation.

A. the average gain registered by a group is independent of the size of the group
B. the efficiency of the teacher is the same for both methods of instruction
C. the relative efficiency of the two instructional methods is dependent upon the nature of the subject matter
D. A and C above
E. none of the above

C7 26. A sample standard deviation is calculated by taking the square root of the quotient of the sum of the squared deviation scores divided by the sample size. This sample standard deviation provides _____estimate of the population standard deviation.

A. a random
B. a biased
C. an unreliable
D. the most powerful
E. none of the above

C8 27. The analysis of variance has the same purpose as

A. the t-test
B. the chi square test
C. the F ratio for sample variances
D. both A and C
E. A, B, and C

P9 28. A 100-point English test was given to all 10th graders in a school over a period of 10 years. Ten per cent of the total group scored above 90. On a history test given to these same 10th graders, the mean score of the 10 per cent who scored highest on the English test would most likely be:

A. closer to the mean of the total group than it was on the English test
B. at X_{90}
C. farther from the mean score of the total group than it was on the English test
D. the same as the mean score for this group on the English test
E. no estimate can be made with any degree of probability.

C9 29. The teacher who understands statistics

A. will not use a test norm as a standard of high achievement
B. considers achievement scores approximate
C. sees causes in correlations
D. will be dismayed if half his class scores below the median for his room
E. both A and B

C10 30. The mean of a distribution can be misleading if

A. all scores are the same
B. the sample is not random
C. one score is many times larger than the sum of the other scores
D. the sum of the deviation scores is not reported
E. the range of scores is more than six times the standard deviation

C11 31. A common misuse of chi-square is to

A. substitute percent for frequency
B. apply the technique to large frequency tables
C. ignore a test of linearity
D. fail to normalize scores before analyzing the data
E. assume the data is only nominal

C12 32. The most common purpose(s) of graphs is (or are) representations of

A. changes over time intervals
B. parameters of a population
C. statistics of a sample
D. relationships among variables
E. both C and D

C13 33. Rounding raw scores _____ the TOTAL error in the scores

A. decreases
B. increases
C. accounts for
D. masks
E. does not affect

C14 34. For a given sample size, the confidence that the parameter is within a given interval will become smaller as the interval becomes:

A. more probable
B. wider
C. less probable
D. narrower
E. more information is needed before anything can be said

C15 35. No matter how long you look, you still won't discover anything about the relationship between two variables by examining

A. the Pearson product moment coefficient of correlation
B. the scatterplot
C. the covariance term
D. the two variances
E. the slope of the regression line.

C16 36. Although the investigator does not know it, in the population boys and girls are equally capable of learning Lesson 14. The probability that any $\underline{t}$ test will result in the conclusion that boys are different from girls in this respect should be indicated by the

A. level of significance
B. efficiency of the statistic
C. power of the $\underline{t}$ test
D. probability of a Type II error
E. both C and D

P10 37. Suppose we multiply each test score, in an approximately normal distribution of test scores by 5, then add 15. The resulting distribution of the new variable would be approximately

A. binomial
B. normal
C. skewed
D. rectangular
E. could be any of the above

F12 38. The standard deviation is an index of

A. covariance
B. homogeneity
C. skewness
D. random errors
E. statistical inference

F13 39. The standard error of estimate is the

A. standard deviation of prediction errors
B. standard deviation of sampling errors
C. reliability coefficient for prediction errors
D. reliability coefficient for sampling errors
E. both C and D

P11 40. In a controlled experiment with 12 subjects in each of two groups a researcher used a " t" test when he could have used a "z" test. What effect did this mistake have?

A. increased the power
B. increased the probability of rejecting a false null hypothesis
C. decreased the probability of type I error
D. increased the probability rejecting a true null hypothesis
E. both A and D

C17 41. In an article discussing error rate in tests of hypotheses, one writer stated that he was primarily concerned with minimizing the error that occurs when we falsely conclude that two population means are different. This error is a

A. sampling error
B. bias
C. type I error
D. type II error
E. both B and D

C18 42. The Spearman rho can be used whenever

A. the same group of individuals is ranked twice with regard to two different attributes
B. two groups of individuals are ranked with regard to a single attribute
C. the same group of individuals is ranked once with regard to a single attribute
D. two groups of individuals are ranked twice with regard to two different attributes
E. both A and B

F14 43. In constructing grouped frequency distributions, which of the following rules of thumb is generally disregarded?

A. the width of the interval should be an odd number
B. the number of intervals should be between 10 and 25
C. the intervals should be of equal width
D. the lowest score should be placed at the middle of the lowest interval
E. score limits rather than real limits should be used to define the interval

P12 44. In his class Johnny's score on test A was at the 95th percentile; his score on test B was also at the 95th percentile. If each student in Johnny's class had his scores from test A and test B added together, at what percentile would Johnny's combined score PROBABLY be amongst all the combined scores, IF TEST A AND B WERE CORRELATED -1.0?

A. less than 95P
B. equal to 95P
C. greater than 95P
D. such a condition could not exist
E. more information is needed before any reasonable guess can be made

P13 45. As in the previous question, Johnny's score on test A was at the 95P, as well as on test B. Again scores were combined. At what percentile would Johnny's combined score PROBABLY be amongst all combined scores IF TEST A and B WERE CORRELATED +0.4?

A. less than 95P
B. equal to 95P
C. greater than 95P
D. such a condition could not exist
E. more information is needed before any reasonable guess could be made

APPENDIX B

Essay Tests and Keys

Due to the fact that doctors did not know, prior to 1960, that direct optical exposure to oxygen could cause blindness, some premature babies born prior to that time lost their sight when placed in oxygen tents. In Manitoba and Alberta alone there are about 200 such cases. Such children in Manitoba now attend a blind school in Winnipeg; Alberta's unfortunates attend a school in Edmonton.

A valid instrument for measuring reading aptitude amongst the blind was administered to Manitoba's and Alberta's children mentioned above. Two researchers, both using this collected data, did some statistical analyses. The first researcher (R-1) did a t-test for differences of means and obtained the following results:

	Mean	df	t	p(2-tail)
Alberta	99			
Manitoba	101	198	0.94	0.34

From this analysis R-1 concluded the mean reading aptitude for Alberta's oxygen-blinded children was equivalent to that of Manitoba's oxygen-blinded children..

The second researcher (R-2) simply calculated the same means and concluded the means were different, but not meaningfully different.

You are requested to COMPARE and EVALUATE R-1's and R-2's analyses and conclusions in terms of the following:

i) hypotheses made by R-1 and R-2
ii) statistical techniques and their appropriateness
iii) sample size and sampling error
iv) probability of correct decisions
v) probability of incorrect decisions
vi) which of R-1 or R-2 results tells precisely the differences which really exist.
vii) advantages and disadvantages of using each analysis to determine:

- statistically significant differences
- substantive or meaningful differences

NB- use the vernacular of statistical analysis and explain the meaning of terms and/or symbols specific to statistics.

OXYGEN BLINDED STUDY — KEY FOR ESSAY 1 — ID_____

FACTUAL TRAIT	max	obt
Hypothesis:		
$\mu Man = \mu Alta$ & $\mu Man \neq \mu Alta$	1	
null, alternate (Ho, H_1)	1	
Technique: inferential, descriptive ..	1	
Sample:		
-sample is subgroup of population	1	
-sampling error: unrepresentativeness of population by sample ($S_{\bar{X}}$)	1	
-estimates (statistics), parameters ..	1	
-finite (limited) and infinite population	1	
Correct Decision:		
-if Ho true: $1-\alpha$ (level of confidence)	1	
probability of not rejecting true Ho	1	
-if H_1 true: $1-\beta$ (power of test)	1	
probability of not rejecting a true H_1	1	
Incorrect Decision:		
-if Ho true		
-α (type I error) (level of significance).	1	
-probability of rejecting a true Ho	1	
-if H_1 true		
-β (type II error)	1	
-probability of rejecting a true H_1	1	
TOTAL FACTUAL	15	
COMPREHENSION TRAIT		
Hypothesis: both had same	1	
Technique: -R-1: inferential	1	
-R-2: descriptive	1	
Sample: -sample here is population	2	
-R-1: assumed sampling error .	1	
-R-2: assumed no sampling error	1	
Correct Decision:		
-R-1: probability set $1-\alpha \geq .95$	1	
-tho't if were to say difference, $1-\alpha = .66$	1	
-didn't know his $1-\beta$	1	
-unless value of H_1 difference specified	1	
Incorrect Decision:		
R-1 - tho't $\leq .05$ not attained	1	
- tho't $\alpha = .34$ if were to reject Ho	1	
- committed type I error	1	
- didn't know his β	1	
TOTAL COMPREHENSION	15	

PROBLEM SOLVING TRAIT	max	obt
Hypothesis -R-1 conclusion wrong	1	
-R-2 conclusion correct	1	
Technique -R-1: inferential inappropriate for this study	1	
maybe appropriate to all blind in Alta/Man	1	
but if for all blind, maybe this type type of blind subjects are biased...	1	
R-2 description correct for finding statistically different means in measured finite population.	1	
Decision:		
R-2: $1-\alpha = 1.0$ (perfect) OR $\alpha = 0.0$	1	
$1-\beta = 1.0$ (perfect) OR $\beta = 0$	1	
Precision only parameters give precise real differences, hence descriptive suitable	1	
always uncertain of parameter if inferential	1	
Advantages:		
Description: if have parameter no inference needed for statically significant differences	1	
: allows direct inspection for meaningful differences	1	
Inferential: allows inspection of meaningful differences only if		
-significant differences occur	1	
and -if value of H_1 specified	1	
Neither show meaningfulness - that is determined by criterion independent of statistics	1	
TOTAL PROBLEM SOLVING	15	

B.2: Essay 2

Two researchers (R-1 and R-2) attempted to determine if mean IQ's of specialists ($\bar{X}sp$) and general practioners ($\bar{X}gp$) were different in Alberta's population of doctors. Both used the same IQ test which had a test-retest correlation of .95. Both used a t-test to analyze their data. The results and conclusions of R-1 and R-2 are as follows:

Researcher	IQ test	N used in each sample	$\bar{X}sp$	$\bar{X}gp$	P(2-tail)	Conclusions
R-1	Form A	25	120	124	.06	$\mu sp=\mu gp$
R-2	Form A	200	122	123	.03	$\mu sp\neq\mu gp$

You are requested to COMPARE and EVALUATE the paradoxical results and conclusions of these two studies, in terms of the following:

i) probable hypotheses and assumptions made by R-1 and R-2
ii) statistical techniques and their appropriateness
iii) size of sample and sampling error
iv) probability of correct decisions, if: Ho true, Ho false
v) probability of incorrect decisions, if: Ho true, Ho false
vi) which of R-1 or R-2 results tells precisely the IQ differences which really exist
vii) advantages and disadvantages of using this type of statistical investigation to determine
- statistically significant differences
- substantive or meaningful differences

NB- use the vernacular of statistical analysis and explain the meaning of terms and/or symbols specific to statistics.

IQ Specialists Study — KEY FOR ESSAY 2 — ID _____

FACTUAL TRAIT	max	obt
Hypothesis		
population of Alta specialists = population of Alta gen. prac.	1	
null, alternate (Ho, H_1)	1	
Technique: inferential	1	
Sample:		
-sample is subgroup of population.....	1	
-sampling error: unrepresentativeness of population ($S_{\bar{x}}$).....	1	
-estimates (statistics), parameters.....	1	
-finite (limited) and infinite population	1	
Correct Decision:		
-if Ho true: 1-α (level of confidence)..	1	
probability of not rejecting true Ho ..	1	
-if H_1 true: 1-β (power of test).....	1	
probability of not rejecting true H_1	1	
Incorrect Decision		
-if Ho true		
-α (type I error) (level of significance)....	1	
-probability of rejecting a true Ho.	1	
-if Ho false		
-β (type II error).....	1	
-probability of rejecting a true H_1.	1	
Total Factual	15	

COMPREHENSION TRAIT	max	obt
Hypotheses: both had same	1	
Technique:		
-both used inferential technique.....	1	
-even tho' Alberta's population of doctors is finite at this point in time, can think of it being infinite if extrapolating to future doctors, hence inferences suitable.....	1	
-both assumed random independent samples.....	1	
-both assumed $N(O,\delta_{sp}^2)$, $N(O,\delta_{gp}^2)$.....	1	
$\delta_{sp}^2 = \delta_{gp}^2$	1	
-both should have pooled variances.....	1	
Sample:		
-N↑, sampling error decreased.....	1	
Correct Decision:		
-probably both set 1-α ≥ .95 or ..... probably both set α ≤ .05	1	
-N does not affect 1-α (or α).....	1	
-control of 1-α (or α) is arbitrary decision of researcher.....	1	
-1-β(power) ↑ as N↑.....	1	
-1-β(power) ↑ as 1-α↓ or ..... 1-B(power) ↑ as α↑	1	

COMPREHENSION (CONT'D)	max	obt
Incorrect Decision:		
-as α↑, β↓ (i.e. as type I error increased type II error decreased)	1	
-as N↑, β↓	1	
TOTAL COMPREHENSION	15	

PROBLEM SOLVING TRAIT	max	obt
Hypothesis: R-2's conclusion more likely correct	1	
Decisions (Correct or Incorrect)		
-if Ho true		
(a) R-1's probability of incorrect decision was 0.06,		
or	1	
R-1's probability of correct decision was .94,		
-if Ho False		
-probability of either is unknown with information given.....	1	
-but β less for R-2 or	1	
1-β greater for R-2		
with Ho unknown, the probability of either decision being correct is 1.0	1	
or 0.0.....	1	
neither R-1 or R-2 can say exactly if Ho is true or false.....	1	
4 point vs. 1 pt spread probably not due to unreliability of test	1	
4 point vs. 1pt spread probably occured by chance.....	1	
Advantages:		
inferential techniques prevent statement of difference which occurs by chance.....	1	
if N very large, likely always significant.....	1	
Disadvantages:		
if N's small likely no signif. diff...	1	
if N's large, sign.diff found regardless of value.....	1	
one NEVER knows precisely real differences in nature.....	1	
Meaningfulness:		
behavioral research not meant for finding only significant results	1	
criteria of meaningfulness is independent of statistics, hence statistics only a tool.....	1	
TOTAL PROBLEM SOLVING	15	

APPENDIX C

Simulation Documentation for IBM 1500 System

C O M P U T E R - A S S I S T E D - I N S T R U C T I O N

COURSE DOCUMENTATION

Course Title:	*clark*
Educational Level:	Graduate student
Author:	Clarke Hazlett
Computer:	IBM 1500 System
Language:	Coursewriter II
Date:	January, 1972

Division of Educational Research

The University of Alberta

Edmonton, Canada

Information for the Educator

1. Objectives and Purpose

   Simulation in research design. Evaluate ability in research design - i.e., two programs are a test in one's ability to do statistical analysis and design research projects: the student is allowed some latitude in his approach.

2. Course Content

   Ed. Psy. 502/504 - i.e. applied statistics in educational research.

3. Educational Level

   Graduate students.

4. Source of Course Content

   Various journals, background experience of author, etc.

5. Program Duration

   -0: 15 min.; -1: 1 hr.; -2: 1 hr.

6. Method of Instruction

   No instruction is used.

7. Audio-Visual Requirements

   Audio tapes required -23 messages for each segment.

8. Response Modes

   Light pen, space bar.

9. Student Operating Requirements

   Sign-on only; password SA for segment 1 RD for segment 2.

10. Teacher Supervision

Nil.

11. Evaluative Information

Yields three scores: FACTUAL, COMPREHESION, PROBLEM SOLVING reflecting low to high cognitive skills; validity estimates indicate PROBLEM only valid.

12. Student Reaction

Negative first try, positive second try.

13. Programmer and Author

Clarke Hazlett & Wayne Osbaldeston.

14. Availability

Unrestricted, but permission should be requested of author.

Information for Operator

1. Course Name and Segments

   clark -0, -1, -2

2. Dictionaries and Graphic Sets

   Nil.

3. Functions Called

   0 - sb, 1 - sb, mv, fade
   2 - sb, mv, fade

4. Macros Called

   Nil.

5. Film Reels

   Nil.

6. Audio Tapes

   #17 - master.

7. Execution Time

   0 - 15 min., 1 - 1 hr., 2 - 1 hr.

8. Response Time

   All ep's are defaulted.

9. Pre-Course Instruction Requirements

   Use segment 0 - self explanatory.

10. Student Sign-On Command

Special code word required by student -1: *sa*
-2: *rd*

11. Teacher Supervision

No.

12. Proctor Messages

No.

13. Performance Recordings

Possible.

14. Special Instructions for Operator

Audio tapes only.

Information for Programmer

1. Course Listings and How - Chart

   360 output of Documentation is enclosed.
   360 output of How - Charting program is enclosed.

2. Macros Called

   Nil.

3. Special Programming Features

   Nil.

4. Programmers Name and Date of Documentation

   Clarke Hazlett
   September, 1971

DOCUMENTATION TABLE

Segment	Buffers	Counters & Cntr/Sw.	Switches	Ret. Reg.	Functions	Graphics	Dictionaries
0	0	1	1, 31	———	ꝺꝺ	———	———
1	1	1,4,5,6,7, 8,10,11,12, 13,14,15, 16,17,25	0,1,2,3, 7,8,20,21 22,23,24, 31	1	ꝺꝺ mv fade	———	———
2	1	1,5-8,10-17 24-25	0-3,7,8, 20-24, 31	1	ꝺꝺ mv fade	———	———

APPENDIX D

Face Validity Questionnaire

IDENTIFICATION NUMBER (I.D.)_____

What do you think the three types of examinations measure?

For each of the 15 statements in this questionnaire, please indicate the degree to which you agree or disagree with the statement in terms of each type of examination.

SD means strongly disagree
D means disagree
A means agree
SA means strongly agree.

Eg. Statement #9 reads: Great emphasis was placed on memorization. Assume someone filled his response in the following manner.

	SA	A	D	SD
ESSAY	✓			
M.C.		✓		
SIM.				✓

Such a response would mean:

- he strongly agreed (SA) with this statement in terms of the two essay exams he had written.
- he agreed (A) with the statement in terms of the two multiple choice (M.C.) exams he had written.
- he strongly disagreed (SD) with this statement in terms of the two simulation exams he had written.

1. This type of examination gave one the opportunity to demonstrate his abilities in some important areas of research design and data analysis.

	SA	A	D	SD
ESSAY				
M.C.				
SIM.				

2. Most topics covered were irrelevant to research design and data analysis.

	SA	A	D	SD
ESSAY				
M.C.				
SIM.				

3. The examination was too difficult.

	SA	A	D	SD
ESSAY				
M.C.				
SIM.				

4. Remembering or recognizing was the main information tested.

	SA	A	D	SD
ESSAY				
M.C.				
SIM.				

5. This type of test measured one's skill in recognizing and weighing values in alternative courses of action.

	SA	A	D	SD
ESSAY				
M.C.				
SIM.				

6. One had to actively put methods and ideas to use in new situations.

	SA	A	D	SD
ESSAY				
M.C.				
SIM.				

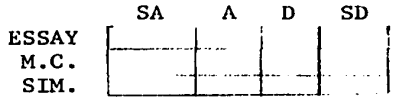
7. One was expected to go beyond the information given to see what was implied.

	SA	A	D	SD
ESSAY				
M.C.				
SIM.				

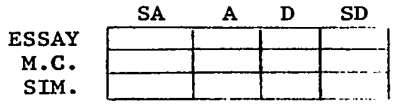
8. Great importance was placed on logical reasoning and analysis.

	SA	A	D	SD
ESSAY				
M.C.				
SIM.				

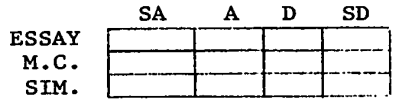
9. Great emphasis was placed on memorization.



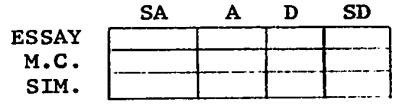
10. One was required to work with pieces, parts, elements, etc. and arrange and combine them in such a way so as to constitute a structure that was not clearly there before.



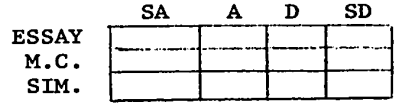
11. Using logic and reasoning processes to think through complicated problems (and prove the answer) was a major activity.



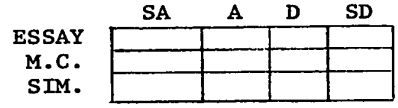
12. A central concern was practising methods in life-like situations to demonstrate skill in solving problems.



13. One was expected to read between the lines to find trends and consequences in what was presented.



14. A major job was to make judgements about the value of issues and ideas.



15. One was tested on his ability to formulate appropriate hypotheses and modify such hypotheses in the light of new factors and considerations.