DOCUMENT RESUME

ED 224 811 TM 820 813

AUTHOR	Hogan, Thomas P.
TITLE	Relationship between Free-Response and Choice-Type Tests of Achievement: A Review of the Literature.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	[81]
NOTE	51p.
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Achievement Tests; Correlation; Essay Tests; *Measurement Techniques; Multiple Choice Tests; *Objective Tests; *Test Format; *Test Selection
IDENTIFIERS	*Free Response Test Items; National Assessment of Educational Progress

ABSTRACT

Do choice-type tests (multiple-choice, true-false, etc.) measure the same abilities or traits as free response (essay, recall, completion, etc.) tests? A large number of studies conducted with several different methodologies and spanning a long period of time have addressed this question. In this review, attention will be focused almost exclusively on the measurement of the traditional product of education, namely knowledge. This review is limited to empirical studies of the equivalence of free-response and choice-type tests. The major methods used to study the relationship between free-response and choice-type measures are the direct correlation, the criterion correlation and the treatment effect. Contrary to widely held beliefs about choice-type tests, the studies indicate that the two types of tests do generally measure the same traits or abilities. To the extent that there are minor differences, the choice-type measures tend to be more valid; and use of choice-type measures does not seem to have adverse effects on study habits. However, the generalizations are limited by insufficient diversity in the groups studied and may not apply to certain types of more divergent processes. Aspect of National Assessment (NAEP) dealt with in this document: Assessment Instrument (Multiple Choice Exercises). (Open Ended Exercises). (Author/PN)

RELATIONSHIP BETWEEN FREE-RESPONSE AND CHOICE-TYPE TESTS OF ACHIEVEMENT:

A Review of the Literature

Thomas P. Hogan

University of Wisconsin-Green Bay

This review was prepared for the National Assessment of Educational Progress, a project of the Education Commission of the States, funded by the National Institute of Education.

Abstract

Do choice-type tests (multiple-choice, true-false, etc.) measure the same abilities or traits as free response (essay, recall, completion, etc.) tests? A large number of studies conducted with several different methodologies and spanning a long period of time have addressed this question. Contrary to widely held beliefs about choice-type tests, the studies indicate that the two types of tests do generally measure the same traits or abilities; to the extent that there are minor differences, the choice-type measures tend to be more valid; and use of choice-type measures does not seem to have adverse effects on study habits. However, the generalizations are limited by insufficient diversity in the groups studied and may not apply to certain types of more divergent processes.

CONTENTS

RELATIONSHIP BETWEEN FREE-RESPONSE AND CHOICE-TYPE TESTS OF ACHIEVEMENT:

A Review of the Literature

THE PROBLEM

To determine with some degree of accuracy and objectivity how much students know (or don't know) is the central problem of educational testing. The determination may be made for a variety of purposes--grading, diagnosis, program evaluation, etc.--thus tilting a particular application in this direction or that, but the basic problem remains the same.

The problem, of course, is not new: It goes hand-in-glove with the educational process. Educators, or their external supervisors, have always "tested" students (DuBois, 1970; Ebel, 1972), although in times long past (say before 1850) testing, as well as instructional methodology, was not open to question. Everyone knew precisely how to do it! We're not so confident today.

For reasons which may some day be divined by cognitive psychologists or evolutionary biologists, there is an overwhelming inclination to believe that the "natural," "correct" or "direct" way to assess student knowledge is to put a question to the student and have him/her respond in a free and open manner. Such responses, referred to in testing jargon variously as free-response, open-ended, or constructed responses, may be given orally or in writing. In the latter mode, they are sometimes referred to generically as "essay" tests, although in some instances the "essay" may be as short as a word, a phrase, or a number.

In contrast to the free-response method of testing, we have the now familiar choice type items, one of the most distinctive contributions of behavioral sciences to contemporary society. In choice type items, the

examinee is presented with a number of alternative answers to the question and chooses, most typically, one of these answers as correct. The most popular forms of choice type items are the multiple-choice variety and the true-false item, which is really just a specific case of a multiple-choice item, i.e. one with two choices (true or false) as possible answers. In addition to the multiple choice and true-false types of items, numerous other choice type items have been devised and experimented with. Much research has been conducted with variations in formats, directions, and scoring producers for this or that choice type item.

The enduring question for the choice type items is whether or not these seemingly artificial contrivances measure the same thing as the more "natural and direct" free-response types of item. Popular opinion on this question is rather well formulated and almost universally negative, i. e. the two types of items do not measure the same thing. One can hear multiple-choice and true-false questions castigated in nearly any teachers' lounge in the country on a daily basis, and they are lampooned with regular frequency in cartoon strips (perhaps the best "social indicator" of the pervasiveness of the choice type item). In addition, professional journals and books in the education field routinely lambaste the alleged triviality, ambiguity, and assorted other evils of the choice type question, not infrequently reaching a feverish pitch.

But at root the question of whether free-response and choice-type tests are measuring the same thing (trait, ability, level of knowledge) is an empirical one, not a philosophical or polemical one. The concepts and methodology to determine the equivalence of the two types of measures have been available for a little over 50 years and have, in fact, been applied in dozens of studies. In this review, we wish to "pull these studies together" to determine to what extent research has provided an answer to the question regarding the equivalence of the two types of measures. In this review, attention will be focused almost exclusively on the measurement of the

traditional product of education, vis. knowledge. No reference at all is made to studies in the realm of affect or personality. Furthermore, we have chosen to exclude the communication skills of reading, writing and speaking, not because these skills fall outside our concern for the products of education, but because they seem to present sufficiently unique cases to warrant coverage in separate reviews. Finally, we limit the review to empirical studies of the equivalence of free-response and choice-type tests; no attempt will be made to review strictly rhetorical analyses of the question.

Before concluding this introductory section, it might be noted that the question under review is of considerably more than academic interest. The question has substantial financial implications. It turns out that in most, though not necessarily all, instances which involve large-scale testing, it is much less expensive to use choice-type items than to use free-response items, due to differences in scoring costs. Although detailed cost comparisons would have to be made within the context of a specific project, it would not be unusual for the cost of an assessment endeavor depending heavily upon free-response measures to cost twice as much as a similar project depending heavily upon choice-type measures or, conversely, for the choice-oriented project to cost half what the free-response project would cost. When one contemplates an assessment project costing, say $500,000, the importance of knowing whether the two types of measures are equivalent becomes poignantly clear.

METHODS OF STUDY

Although the methodologies employed in specific studies will be discussed as each study is introduced in subsequent sections, it will be convenient to outline first the major methods which have been used to study the relationship between free-response and choice-type measures. There appear to be three basic methodologies in use to attack the problem. We shall refer to them as

the direct correlation, the criterion correlation, and the treatment effect methods.

Direct Correlation. In the direct correlation method, the correlation (usually the Pearson) between a free-response and choice-type measure is determined; most frequently, the correlation is corrected for attenuation (see next section). If the corrected correlation approaches a value of 1.00, it is concluded that the two types of tests are measuring the same trait, variable, ability or skill. If the corrected correlation departs substantially from unity, obviously one concludes that the two types of tests are measuring somewhat different things, and the authors usually express a preference for one or the other types of measures based on criteria other than what is being measured. Such other criteria include reliability, breadth of content sampling, efficiency, face validity, examinee preference, effect on students' study habits, as well as many other matters.

The direct correlation approach has been, by far and away, the most frequently used methodology in this area. It is simple to apply and yields data that are relatively easy to interpret. However, interpretation of results from studies using this approach is subject to some personal inclinations. For example, a correlation between free-response and choice-type tests, corrected for attenuation, of .90 in one person's book is high enough to warrant the conclusion that the two tests are measuring the same thing for all practical purposes, while in another's book it is low enough to show that the two tests are not measuring precisely the same thing.

Criterion Correlation. In the criterion correlation approach, both free-response and choice-type tests are correlated with some external criterion which is taken to be, in some sense, a better measure or precisely the measure of the variable of interest. The type of test which yields the higher correlation with the criterion is considered the better measure. Note that in this approach one assumes at the outset that the two types of tests

(free-response and choice-type) are probably measuring somewhat different things, the only question being which yields the better approximation to the external criterion.

The correction for attenuation may also be applied in this approach although the correction is usually made for unreliability in the test only, not in the criterion. Use of the correction for attenuation in this approach may present a potentially thorny problem of interpretaton. Let us say that F represents a free-response measure, C a choice-type measure, and X a criterion; r(FF') = .50, r(CC') = .90, r(XF') = .40, and r(XC) = .50. At this point, the choice type test is better because it correlates more highly with the criterion. However, when the correlations between the tests and the criterion are corrected for unreliability in the tests, the free-response test becomes better, i.e. correlates more highly with the criterion (r'(XF) = .57, r'(XC) = .53). But is it reasonable to suppose that the free-response test can be made substantially more reliable than it already is? Although there is no simple solution to this type of problem, we should at least be aware of the difficulty as we review studies of this type.

A special problem encountered in the criterion correlation approach is that of "criterion contamination" in which one of the measures being investigated (free-response or choice-type) directly or indirectly affects status on the criterion variable. (The problem of criterion contamination has long been discussed in clinical research on test validity. See Anastasi (1976) for a general treatment of the topic.) When the contamination is direct, e.g. when the free-response measure being studied is an essay test used for a final exam which will contribute 50% to the final grade which will serve as the criterion, then the problem is usually recognized, although rarely does the author attempt to disentangle the effects of the contamination. Potentially more hazardous for clear interpretation of results is indirect contamination in which certain irrelevant sources of variance

affect status on both the criterion and one of the measures being investigated, even though the latter does not directly enter into the criterion. For example, final grade in a course may be determined by 10 quizzes, all of the short answer essay variety; then, at the end of the course and not entering into the determination of the grade, we obtain an essay and a multiple-choice measure of knowledge of course content and correlate scores on these measures with final grade. If we are willing to grant that the criterion itself is not perfect, i.e. not the best possible measure of knowledge of course content because it is affected to some extent by abilities peculiar to taking essay tests (e.g. ability to bluff, snow, etc.) which, in turn, also influence status on the essay test being investigated, then we have a case of indirect criterion contamination. Obviously, it is quite difficult to unravel the influence of all such indirect contaminants but, again, we should at least be aware of this problem.

Treatment Effect. A third possible methodology is to apply some treatment to a group, the effect of which should be to increase scores on some trait or ability, then determine which of several measures is most sensitive in detecting the intended change. The measure which detects the largest change is considered the best. Use of this technique prescinds entirely from any assumption about psychometric equivalence of the measures being studied, but does have considerable intuitive appeal in educational contexts since education may be thought of as the application of a treatment.

Let us provide a practical example. A group of 200 students is divided by random means into two subgroups of 100 each. One subgroup, call it the treatment group, is taught economics one hour per day for two weeks at the end of which both groups take an essay test and a multiple-choice test on economics. Which test better distinguishes the treatment group from the control group? Or are the two tests equally sensitive to the group difference? Quite obviously, the two tests could be measuring very different

things and still show equal differences between the groups or differences favoring either one or the other type of test.

Actually this third methodology could be considered as a special case of the criterion correlation method in which the criterion is considered a dichotomous variable (treatment = 1, control = 0), with results being expressed as a biserial correlation. However, it will be more convenient to treat the two methods separately. It might be noted that the treatment methodology clearly skirts the issue of criterion contamination: assignment to the treatment group in no way depends on any test-taking ability either directly or indirectly.

Given the usual uses of educational tests, the "treatment effect" methodology, while lacking psychometric precision, has a certain intuitive appeal. It is surprising, therefore, that it has been used in only a few studies.

Other Methodological Issues. Although the great majority of studies to be considered later employ one of the three methods just described, there are, as one might expect, a number of other methodological issues of a general nature which merit comment. First and foremost, it should be noted that while each of the three basic methods just described has a simple, direct relevance to the problem in question, they all seem to lack any high-powered, theoretical underpinnings. This difficulty has been attacked rather recently by Lord with his discussions of T-equivalent measurements (Lord, 1971; Lord and Novick, 1968). The exposition does not yet seem complete, but a few studies employing basic notions from this line of reasoning have appeared, e.g. Traub and Fisher (1977) and Ward, et al (1980). There appears to be much unfinished work in this area. Perhaps of special importance is the investigation of equivalence of measures for various subpopulations, since by definition T-equivalent measures must show equivalence within all subpopulations (Lord and Novick, 1968). Only one study could be identified which treated this issue directly

(Longstreth, 1978) and one which treated it indirectly (Peters and Martz, 1931). An appallingly large proportion of the research on the relationship between free-response and choice-type tests is based on the proverbial "college sophomores in general psychology class" or their look-alikes, with not the least hint that conclusions from the study might be limited to this rather unusual segment of humanity. Educational researchers have apparently never outgrown their belief in the infinite generalizability of results from such subgroups: One of the most recent studies in our subject domain (Gay, 1980) is based on two groups of 14 students each in one introductory educational research course.

Finally, we note that on the issue of the effect of different testing methodologies on students' study habits and student preference for different types of tests, the evidence is generally "soft," being based mostly on students' self-reports. However, a few studies have attempted to investigate these matters with more sophisticated techniques (Sax and Collet, 1968; Gay, 1980).

THE CORRECTION FOR ATTENUATION

The correlation between two measures, r(XY), is limited, lessened, or "attenuated" by the imperfect reliability of the two measures, r(XX') and r(YY'). If reliabilities of the two measures are known, the correlation between the measures may be "corrected for attenuation" or "corrected for unreliability" by use of the formula $r(XY)/\sqrt{r(XX')\ r(YY')}$, yielding an estimate of the correlation between true scores on X and Y. As Lord and Novick (1968, p. 69) point out, the attenuation problem is of "fundamental importance" and ". . . is one which first motivated the development of test theory as a distinct discipline."

As might be expected, the correction for attenuation is used prolifically in the literature on the relationship between free-response and choice-type

measures. Indeed, it was precisely this problem which provided one of the early applications for the attenuation formula, as well as for correlation methodology in general; many authors of the 1920 era applied the correction with the delight of a new-found toy, although some authors then and even today seemed oblivious to the formula itself as well as to the underlying notion that imperfect reliability of a measure affects its relationship with other measures.

It is assumed that readers of this paper have some familiarity with the theoretical justification for and application of the correction for attenuation; hence, we do not intend to provide a review of the issue here, beyond the brief outline given above and to note below two special problems in the application of the correction. Readers wishing more information about the correction are referred to such standard sources as Guilford (1954), Gullikson (1950), Lord and Novick (1968), and Stanley (1971).

Applications of the correction for attenuation often run afoul of one basic assumption underlying the rationale for the correction, vis. that determination of r(XY) and the two reliability coefficients, r(XX') and r(YY') are subject to the same sources of error variance. Specifically, it often happens that the reliability coefficients are affected by fewer sources of unreliability than is r(XY). For example, the X and Y measures, which in the particular application of interest in this paper ordinarily represent a free-response measure and a choice-type measure, not only differ in format but also are usually obtained at different times (say two weeks apart) and sometimes under different motivational circumstances (one of the measures being a real final exam, the other being an experimental measure). Hence, a variety of sources of variation may be affecting r(XY), in addition to the difference in test formats. In contrast, for purposes of applying the correction for attenuation in the latter situation, the study's author may calculate the odd-even reliability of the choice type measure and the scorer

reliability of the free-response measure, the resulting reliability coefficients being (probably) substantially higher than appropriate for use in the attenuation formula and the resulting corrected r(XY) being lower than it ought to be.

To comment extensively on the appropriate application of the correction for attenuation for each study taken up in later sections would be unwieldy, but the reader should at least be forewarned of the problem; and we will comment on some apparently egregious misapplications of the formula.

A second problem to which we should be alert is caused by attempts to equate the X and Y measures in terms of content, thus avoiding specific content as one source of variation. In a surprising number of studies, this problem is "solved" by using precisely the same questions (test items) in free-response and choice-type formats. Invariably, the free-response format is presented first, followed by one or more choice-type formats, all using the same item stems. While these item stems are physically the same when they are seen by examinees for a second, third, or fourth time, it is difficult to believe that the stems are the same from time to time in terms of the psychological and experiential make-up of the examinees. When one has been asked four times in succession over a period of several weeks, "In what year did Columbus discover America?" is the only difference in content the fact that the question is followed by a fill-in blank the first time and three choices of dates the fourth time? Only one study which used this type of design commented on the odd situation in which examinees are placed: Traub and Fisher (1977, p. 360) note ". . . the difficulty that is encountered in sustaining student motivation when tests are administered repeatedly" but they neglect to speculate about how this obvious problem--which contributed to reducing their number of useable cases by half--might have affected their conclusions.

We might note that both of the problems just reviewed (undercorrecting for

attenuation and repeated use of identical content with an unknown effect upon examinees) would tend to reduce the reported degree of correspondence between free-response and choice-type measures. With respect to the first problem, it is sometimes possible, by reference to information from outside a particular study, to make an intelligent guess about the magnitude of the undercorrection. For example, multiple-choice tests (in the cognitive domain) with split-half reliabilities of .90 usually have alternate form reliabilities of about .85; so, if a split-half reliability was used in applying the correction for attenuation, and it seems more appropriate to use an alternate form reliability, it may be possible to recalculate the correction. With respect to the second problem, we do not know of any method to estimate its effect; furthermore, its effect may vary substantially from one study to another. Contrasted with these latter problems besetting the interpretation of a disattenuated correlation (r'), discussions of the theoretically most defensible method for testing whether r' differs significantly from 1.00 (e.g. Forsyth and Feldt, 1970; Lord, 1957; McNemar, 1958) seem to pale by comparison. In the vast majority of instances, it seems r' can be interpreted only in a rather rough-and-tumble fashion.

THE EARLY YEARS

Approximately two-thirds (some 40 studies) of all published research on the equivalence of free-response and choice-type measures was conducted in the 1920's and 1930's. By the end of this era, there were some well-formulated generalizations about the equivalence of the two types of measures; these generalizations were passed along in the textbooks on tests and measurements, but gradually references to the original studies began to cease so that eventually--even up to the present--textbook recommendations began to sound more like doctrinaire nostrums rather than empirically based conclusions.

The 1940's, '50's, and '60's saw relatively few published studies on the

relationship between free-response and choice-type tests. The 1970's, especially the latter part of the decade, and on into the early '80's witnessed a renewed interest in the issue, with some new twists to both the formulation of the question and the methodology employed for investigating it. In addition, recent years have seen the emergence of machine scoring technology, which has provided a somewhat novel flavor to the topic, while leaving its substance unaffected. In the light of this rather peculiar historical pattern of research on our topic, it will be convenient to divide the studies into those of the early years (the 20's and 30's) and the later years (1940 to the present). And it will be helpful to introduce the studies from the early years with a brief historical sketch.

Historical Background for Early Studies. Retracing the history of investigations on the relationship between free-response and choice-type measures involves a nostalgic trip through a nether world of education in which professors "regarded" (scored, graded) their students' papers; instructors, in addition to asking the most vaguely worded essay questions, could fire off two hundred items the likes of "Name all the states which border on Kansas" and "What mythological beauty was the cause of the Trojan war?" with nary a twinge of conscience about slighting higher mental processes; and woe betide the student who didn't correctly answer 80% of such questions when his paper was regarded.

What, in this long-gone world, motivated the emergence of what was then called the "new type" test, i.e. the choice-type test? It is supposed by many that choice-type tests are the product of machine scoring demands. Nothing could be more absurd: Machine scoring, even in its most primitive state, did not even exist until the mid-1930's and did not become widely available until the mid-1950's (Baker, 1971; DuBois, 1968). Others suppose that the new type test was developed as a result of some passion to engage in mass, large-scale testing. Equally wrong. It is, of course, true that development of the first

objectively scored intelligence tests, i.e. the various Otis tests, was motivated by a need and/or desire for mass testing (DuBois, 1968; Robertson, 1972). However, the early literature on educational testing is remarkably free from any reference to the need or desire for mass testing. Indeed, even references to the work of Otis were rather rare in the literature on educational uses of the new type achievement test although actual use of the Otis tests, it was apparent from these same sources, was widespread.

The key concern in early investigations of the relationship between free-response and choice-type tests was that of reliability. Odell (1928, pp. 5-6) states the concern succinctly:

> Undoubtedly the chief cause contributing to raise the question [the best form of examinations] was the publication of the results from a number of investigations which showed, or appeared to show, great unreliabiity and variability of the marks given examination papers by teachers. Prominent in making the studies referred to were Johnson, Starch and Elliott, Kelley, and Dearborn. Their work, and also that of others along this line, is too well known and has produced too similar results to justify detailed accounts of the various studies here.

Odell goes on to illustrate some results from the Starch and Elliott reports, which appeared in 1912 and 1913 (without ever giving exact references). This practice of referring to the "well-known fact that ..." or "the many studies showing that . . ." traditional, free-response examinations were unreliably scored without citing particular studies is rampant in the literature of the day. We may assume that the problem was widely discussed in professional meetings and the informal literature of that period.

It is important to note in this connection that the earliest studies of interest in our review were not concerned primarily with the relationship between free-response and choice-type items. Rather, these studies aimed to show that the new type item yielded scores which were much more reliable than scores from traditional examinations. In addition, there was much concern about the relative difficulty of various new-type items (e.g. true-false vs. 3-, 4-, or 5-option multiple-choice, or multiple-choice corrected and

uncorrected for guessing) and about how many items of a given type could be answered by examinees in a certain period of time. Today, it may seem trite or trivial to observe that students get higher scores on a true-false test than on a multiple-choice test using a 5-option format and employing the same item stems. However, this and similar topics were much investigated in the early years. Reports of correlations between free-response and choice-type measures were almost of incidental interest in those studies. And, if those correlations were only moderate, the authors were likely to opt for use of the choice type measure because of its greater reliability and more extensive content coverage.

But the interest in the new type test did not spring forth as a novelty in 1921, the date of the first published study to be considered in our review, nor with the Starch and Elliott revelations of 1912. The issue had been fermenting for at least 50 years. Odell (1928), again, notes that Horace Mann had used the term "new" method of examining at least as early as 1845 in praising the use of uniform written exams in place of oral exams. Ruch and Stoddard (1927, p. 2) note:

> . . . we recall the storm of protest which greeted the statement of Dr. J. M. Rice thirty years ago, before the National Education Association, that the efficiency of the teaching of spelling could be measured by giving the pupils lists of related words to be spelled. . . . Rice is probably entitled to the credit of having produced the first educational test, for as early as 1894-95 he constructed two spelling 'tests' . . . Later he made similar tests in arithmetic and language . . .

DuBois (1968) also notes the nearly evangelistic role played by Rice in sensitizing educators to the gross inadequacies of then-current examination practices.

It should be noted at this point that early references to the "new type" test included practically anything that increased the objectivity of the scoring operation. Hence, the completion or fill-in-the-blank type of item was then considered an objective or new type item, whereas today (and in this

review) we would ordinarily consider such an item a free-response or constructed-response type of item. In 1920, the completion type item was novel, a far cry from the free-wheeling essay in widespread use at the time. Now, because the completion item requires something other than marking a little bubble which will be electronically scanned, this type of item is considered less than fully objective. Interestingly, with further developments in electronic character recognition, the completion item may some day be reconverted to its original categorization with multiple-choice and true-false items.

An Overview of the Classical References. Although there are approximately 40 studies to be reviewed from the 1920-1940 era, there are about a half dozen references which are cited repeatedly in the literature and hence serve as the core or classical references in the field. Without reviewing the actual results of these studies at this point, it may be helpful to sketch briefly the character of these classical references.

One of the earliest and most cited references is the work by Toops (1921): Trade Tests in Education. In this curious little book (118 pages), the author argues in favor of the use of objective exams for measuring vocational skills, e.g. for typists and bricklayers; he criticizes the usual methods for evaluating these skills (basically, an informal interview) primarily on the grounds of unreliability. The book gives over much space to the presentation of actual objective tests for many vocational areas and provides much reliability data. Only a relatively small portion of the book directly addresses the question of the equivalence of free-response and choice-type items.

Two works by Ben Wood were crucial in the early investigations. The first was actually a textbook: Measurement in Higher Education (Wood, 1923). Approximately the first half of this book is, indeed, textbookish in character, providing an historical review of measurement issues and giving

general principles of good measurement practice. However, the last half of the book reads more like an extended series of research reports on the use of the "new type examination" in a wide variety of fields at Columbia College. Extensive data are reported on the relationships between the new and old (essay) exams in such fields as Contemporary Civilization, Physics, Government, Zoology, Economics, Philosophy, Greek Art, etc.

The second work by Wood (1927) appeared in a now seemingly obscure publication of the American and Canadian Committees on Modern Languages. Despite the odd origin, the work itself was monumental, involving, for example, at various times in 1925 and 1926, all of the junior high school students taking French or Spanish in New York City and students taking French, Spanish, German and Physics exams in the Regents program throughout New York state.

Two of the most widely cited works in favor of the use of the new exams were textbooks which contained no new data on the question of interest in this review, nor even reasonable summaries of the studies then available. These were the texts by McCall (1922) and Odell (1928). Later authors cited these two works as if they firmly established the superiority of the new type exams. In pleasant contrast to the aridity of the latter works, we have the oft-cited texts by Ruch (1924, 1929) and Ruch and Stoddard (1927) which provide rich summaries of available studies, and even, on occasion, new data not published elsewhere. Ruch himself was author of a number of the best studies of this era and published with Rice (G. A., not the J. M. Rice referred to earlier) a most remarkable collection of objective examinations collected in a national competition for which monetary prizes were awarded for the best exams submitted in a dozen different curricular fields (Ruch and Rice, 1930).

Finally, we note the curious "study" attributed to Hawkes, then Dean of Columbia College. Actually Hawkes did not author any study on the new type

examination. Rather, a news article in School and Society (1922, p. 141-142) reports that Dean Hawkes claimed that " . . . accuracy of the grades . . . has increased from 65 percent . . . to 90 percent . . . [and] . . . the new examination renders the final grade nearly fifty percent more accurate." Reference is probably being made to Wood's (1922) work mentioned above. What is meant by the statement attributed to Hawkes, unless it refers simply to reliability of scores, is not at all clear and, in any case, hardly represents an empirical study.

Major Studies of the Early Years: Direct Correlation Studies. We will divide the major studies reported in the 1910-1940 era according to the methods of study outlined in a preceding section and conclude with a review of several summaries written during the period.

The most frequent type of study involved the direct correlation method; twelve such studies, some of which actually included a number of substudies, were identified. The earliest study was reported by Toops (1921) who administered three forms of a 50 item "general information test" to 124 students at Teachers College (Columbia). The three forms were recall (completion), true-false, and five-option multiple-choice, administered to different subgroups in varying orders, and having exactly the same item stems. The items were highly factual in nature. Toops reports the average correlation of each test with the other two tests; although these correlations are not corrected for attenuation, it is possible to calculate estimates of the corrected r's from other data provided in the article and these corrected r's all turn out to be slightly in excess of 1.00.

Wood's (1923, 1927) studies are so extensive as to prohibit detailed presentation. As indicated previously, the first work involved use of new and old (traditional, lengthy essays) in numerous courses at Columbia College. The 1927 work involved thousands of students at the junior high school level in foreign language courses (and to a small extent in physics courses) over

several years and in many schools. Much of this work relates to topics not of major concern in this review, e.g. distributions of scores, relations between part scores, etc. However, concerning our topic, Wood's conclusion is virtually unvarying from one set of data to another: ". . . the new examination measures academic achievement as completely as the essay examination, and with greater reliability" (Wood, 1923, p. 207). The consistent theme running through Wood's writings is that the new test correlates as highly with the old exam as the old exams correlate among themselves.

Paterson (1926, p. 247) gives one of the earliest explicit formulations of the contemporary concern that essay tests tap higher mental processes while objective tests measure rote learning of facts.

> By hypothesis, if the two types of examinations are measuring radically different mental functions ('reasoning' as opposed to mere 'information') then their intercorrelations should be considerably lower than the reliability coefficients of either. On the other hand, if the old type examination correlates as closely with the new type examination as it does with itself, then the two are measuring the same mental functions and the asserted difference becomes a verbal difference without any existence in fact.

Paterson's work actually involves a series of studies conducted over a two year period with students in an "orientation course" (no further description) at the University of Minnesota. N's for different sets of data reported in the article range from 68-201, with an average of about 100. Throughout the courses old type and new type tests were used. Specific descriptions of what is meant by new and old type exams are lacking. Twelve intercorrelations between new and old type exams are given (for various terms, finals, midterms, etc.), the average r being .52. Average reliability coefficients (actually intercorrelations within type of test) are given as .67 and .52 for new and old type tests, respectively. Applying the correction for attenuation using these average figures, which the author does not do although he seems familiar with the underlying concept, yields a corrected correlation between new and

old type test of .88. Paterson (p. 248) concludes: ". . . there seems to be no escape from the conclusion that the two types of examination are measuring identical things. Such a conclusion, of course, holds only for the examinations so far used in this course." (A rarely encountered concession that this may not be the last word on the subject.) The Paterson work seems especially noteworthy for its realism, in comparison to some of the contrived or very limited situations encountered in other studies.

Ruch and Stoddard (1927) report an investigation similar in design to that of Toops'. Correlations were obtained between a recall (completion) test and multiple-choice and true-false versions of the same items. The basic questions in the study dealt with corrections for guessing and instructions regarding guessing, but these matters need not concern us here. Data were collected from 2453 students in grades 7, 8, 11, and 12, a refreshing departure from educational psychology classes. Test items, as in the Toops study, were highly factual in nature. Average correlations between the recall test and several different choice-type tests (corrected for attenuation with data provided incidentally in the report) were all in the .90's.

Corey (1930) administered a three hour final examination divided about equally between new type items (96 multiple-choice and 13 matching exercises) and old-type items (6 essays) to 102 students in an educational psychology class. The essays were scored "as objectively as possible." The correlation between "new" and "old" parts of the exam, corrected for attenuation, was .93, prompting Corey to conclude ". . . that new type and essay examinations measure very nearly the same thing . . ." (p. 849).

Laird (1923) reports the remarkably low correlation of .038 between an essay exam and an objective type exam, the latter actually consisting of completion items and, therefore, technically not meeting our definition of an objective exam. The study was based on 54 students in an elementary psychology class. What sense can be made of this remarkably low correlation

between the two exams? A careful reading of the Laird article suggests that he did not actually correlate scores on the two tests in the usual manner. Note the following paragraphs from Laird (pp. 123-124), with our underlining added:

> The results are startling. By grading the essays on the basis of distribution of merit a high correlation might have been obtained between these two examinations. But when they are compared in the amount of information contained the correlation becomes low and the differences very great.
>
> The correlation . . . is +0.038 . . . In the essay examination the students knew approximately half as much as on the other, when checked off against the points that they had been given in the course of the subject under examination. A correlation on the basis of grades given might have been high. But should the sparse sampling of the essay test be considered a true measure of the student's knowledge when other tests show that he knows really twice as much about the subject as he has written?

Obviously, Laird's real concern is that students display less familiarity with the subject matter when tested with a single, wide-open essay than when tested with specific, objective questions. What the correlation reported by Laird actually represents is not at all clear. In general, the study should probably be dismissed as meaningless.

Hurd (1932) correlated scores on "short answer" and multiple-choice forms of a test designed to cover the same content. The report is curiously free of details about the study. No mention is made of the subject matter of the exams, although one infers it was physics since one sample item concerns voltage and a footnote elsewhere refers to a physics workbook; there is no mention of the age or grade level of students involved; and no reference is made to the number of test items. However, it is not difficult to infer from the tone of the article that the author is not terribly enchanted with the new type test. In any case, Hurd reports a correlation of .78 between the two tests which, when corrected for attenuation, becomes .89. It is concluded (p. 29) ". . . that the two tests do not measure exactly the same functions." Hurd goes on to state: "In other words, both are not equally valid in measuring the planned outcomes of the instructional unit," an entirely

gratuitous assumption, unless one posits on some a priori grounds that one of the tests is more valid than the other.

As a separate analysis, Hurd reports that the short answer test correlates more highly with a third test than does the multiple-choice test (data based on "a few schools") and concludes that this points to the higher validity of the short answer test. However, there is no description of this third test other than that "[It] is a form appearing in a work and test book by the writer." Finally, Hurd notes that students showed a greater average raw score gain from "preliminary to final test" on the short answer test than on the multiple-choice test. However, standard deviations are not presented, so one cannot determine whether the difference is simply attributable to scale effects. Hence, the Hurd study, while appearing to show the superiority of the free-response type of test, produces only one unambiguous result, vis. a correlation (corrected) of .89 between the two tests, which by most people's standards is high enough to conclude that the two tests are measuring very nearly the same thing.

In an article remarkably similar in tone to that of Hurd's, Magill (1934) gave three forms of a 50 item "miscellaneous information" test--recall, multiple-choice, and true-false, in that order and in immediate succession--to two of his classes (N's of 41 and 54). In one of the classes, time limits of 7, 5, and 3 minutes (!) were used for the tests, while no time limits were used in the other class. The author notes that few students finished the test under the timed conditions. Correlations between recall and multiple-choice forms were .88 and .91 in the two classes. These are uncorrected for attenuation and the article gives no reliability data which would allow calculation of the correction. However, it seems safe to estimate that the corrected correlations would approach unity. Correlations between recall and true-false forms were .61 and .76 and between true-false and multiple-choice .60 and .91 for the two classes. Magill notes that the "intercorrelations are

in general high, and they are of the order of size of those reported by Ruch [1929]," the latter reference to be reviewed momentarily. However, Magill goes on to discuss, at length, inconsistencies in how students answered particular questions from one form of the test to another and concludes on this basis that the forms measure different "mental functions." Magill does not seem to understand the concept of test validity or equivalent measures.

Weidemann and Newens (1933) administered an "improved compare-and-contrast essay test" followed by an 80 item true-false test "covering approximately the same content of instruction" in each of four sections of an introductory education course, with the procedure being repeated seven times throughout the semester. Median reliability for the essay tests was .69 and also .69 for the true-false tests. The median intercorrelation between essay and true-false tests was .47, which would yield an r = .68 when corrected for attenuation. This figure is low in comparison with those obtained in many similar studies.

Eurich (1931) constructed four test forms (essay, completion, multiple-choice, and true-false) for each of two courses, one in statistical methods and one in educational psychology. The essay exam was constructed first, with the other exams tailored to cover the same content; also the essay was designed to be "somewhat more objective than is usually the case," being scored according to what we would now call a point system. We have here an interesting case where there are two forms of free-response tests (essay and completion) and two forms of choice-type tests (multiple-choice and true-false). Intercorrelations across forms tended to be about the same as within forms; for the educational psychology class all correlations between types of tests, corrected for attenuation, tended to approach unity, while the corrected r's for the statistics class were rather uniformly lower (.62-.81). Eurich (p. 277) concludes:

> The intercorrelations of the tests in the course in educational psychology suggest that if reliable tests are constructed, one of the four types used is probably as adequate as any of the other three for measuring the amount of information which the members of the class

> have accumulated. The evidence for this suggestion is not as clearcut in the class studying statistical methods.

[and]

> If the composite score on three types of examinations is used as the criterion for estimating the validity of the fourth type, the results indicate that the four types of tests have approximately equal validity.

Eurich also notes that students express a preference for the multiple-choice and true-false types of tests.

Carter and Crone (1940) report a study in which the primary concern is the effect on reliability of using certain procedures for revising first drafts of tests, but in the process of investigating this effect correlations between true-false, multiple-choice, and completion items are obtained. Data are based on 125-143 students in an educational psychology course; the exams were mid-terms and finals, each of which contained the three types of items. Carter and Crone (p. 367) conclude: "The inter-correlations between true-false, multiple-choice, and completion tests covering the entire course tend to be high. The range of such correlations is small. The average value is .68 for raw correlations and .84 for correlations corrected for attenuation." Inspection of original tables in the article indicates that correlations involving the true-false tests consistently ran lower than for the other tests; correlations (corrected) between multiple-choice and completion items were .88-.90.

... Criterion Correlation Studies. We now turn to a number of studies which concentrate on the correlations between some external criterion and, respectively, free-response and choice-type tests. The most extensive investigation of this type was conducted by Peters and Martz (1931). For an entire school system (but a small one--120 high school students, 132 elementary students), in all subject areas, throughout grades 3-12, for eight testing periods spread over the school year, five types of examinations were used: true-false, four option multiple-choice, completion and essay. Scores

on these tests were then correlated with teacher-assigned final marks. Results are reported separately by grade level and curriculum area, resulting in a plethora of data. Across all grade levels and areas, the multiple-choice, completion, and essay tests showed about the same degree of correlation with final marks, with true-false tests being somewhat lower. The authors comment on some differences favoring one or the other type of test at the elementary or high school levels, but the differences are generally slight, of the order of .02 (e.g. .76 vs. .78). The first conclusion given in the report seems to carry the main message: "The different tests do not vary greatly in their validity." The validity coefficients tend to be quite high, averaging about .75.

Two critical difficulties beset interpretation of the Peters and Martz study. First, there is little description of the various tests used, e.g. their quality, reliabilities, content coverage, etc. Second, there is no way to determine to what extent teacher-assigned final marks may have been influenced by one or more of the tests (the criterion contamination problem). If we assume that criterion contamination was not a significant problem, that the various tests were approximately equivalent in reliability and content coverage, that the reliabilities of the various tests were about .80 and the reliabilities of teacher-assigned marks about .60 (a typical figure encountered in other studies), then one would conclude that for all practical purposes the various tests used in this study were measuring the same thing.

Gates (1921) investigated the correlation of essay tests and true-false tests with an Achievement Criterion, made up of the sum of the essay tests, the sum of the true-false tests, a grade for homework assignments and a grade for class recitations, oral quizzes, etc. Data were based on students in 10 educational psychology classes at Columbia Teachers College. Individual true-false tests showed an average correlation of .65 with the criterion, while individual essay tests correlated .56 with the criterion. No

reliability data are given for the tests which would allow correcting the r's for attenuation. Gates concludes (strongly) in favor of the use of true-false tests in place of essays.

Knight (1922) correlated scores on a 43 item true-false test with final grade in a college physics course and contrasted this correlation with the correlation between a traditional final exam and final grade (with effect of final exam removed), leaving the final grade affected by three (earlier) exams, frequent short quizzes, and lab work. The effect of indirect criterion contamination in this study cannot be estimated, although it would appear to have some influence. The correlation of final grade with final exam was .64 and with the true-false items .58. Knight notes that the true-false test (only 11 minutes) perhaps could have been longer. In any case, he concludes favorably about the validity of the true-false test ". . . to substantiate Gates' faith in the usefulness of [this] technique in college testing (p. 79)."

It should be noted that Wood's (1923, 1927) studies reviewed above with the other "direct correlation" studies also included considerable data on the relative correlation of new and old type tests with external criteria. These data simply confirmed Wood's conclusions that the objective and essay tests had approximately equal validity.

... Treatment Effect Studies. Two studies in the early years investigated the sensitivity of free-response and choice-type tests to some kind of educational treatment. Crawford and Shulson (1925) established three groups of students (all college): group A was allowed time for thorough study of material on a variety of topics, group B was allowed time only to skim the material, and group C was given no opportunity to study the material. Both essay and true-false tests based on the material were then administered to the three groups. Curiously, the authors comment that they think the true-false tests may not have been very adequately developed. In any case, the two types

of tests distinguished among the three treatment groups with about equal sensitivity.

In a similar study, Shulson and Crawford (1928) administered a 20-item true-false test and a 20-item completion test to two groups of students, one of which had just studied some assigned material and the other of which had not studied this material. The procedure was replicated with eight groups, six at the college level and two at the high school level. The authors conclude (page 583): "The most important outcome . . . [was that] of the total eight experiments, four were favorable to the true-false and four to the completion test. In other words . . . the two tests are equally good, and equally able to distinguish the students who have studied the lesson from those who have not." The authors note the problems encountered in scoring the completion items and conclude generally in favor of the true-false test.

Other Studies. In addition to the studies reviewed under the three major types of methodology, there are about a dozen studies published in the '20's and '30's which are just tangentially related to our topic, or are so poorly designed (or reported) as to defy reasonable interpretation, or they have some minor quirk preventing their easy classification with studies reviewed earlier. However, for the sake of completeness, we will comment briefly on these studies.

Andrew and Bird (1938) compared recall and recognition type items answered by college students, but limited their attention to the relative difficulty of the items; no correlations between the two types of tests were reported. Arnold (1927) was also concerned mainly with the relative difficulty of true-false and recall tests, but did report correlations. However, the tests contained strangely concocted "ridiculous" statements; and the reporting of the relationship between tests is so bizarre as to suggest that attempting to make sense of the report would be futile.

Guiler (1929), too, while referring to "validation" of spelling tests

concentrates exclusively on the relative difficulty issue. He compares written recall, oral recall, and multiple-choice methods. Proceeding on the sadistic assumption that the test which yields the lowest score is the most valid, he concludes that the multiple-choice test is the least valid!

Bayles and Bedell (1931) report correlations between several types of tests, some of which appear at first glance to be free-response but, upon closer inspection, all of the tests are of the choice-type. Bird and Andrew (1937) provide an extensive report on the "comparative validity" of several item types. If one were to read only the conclusions from this report, it might be assumed that tests with various item types were correlated with some external criterion. However, it appears that the authors simply summarized item discrimination indices within each type of test item, rendering the study irrelevant for our purpose (and perhaps for any other purpose, too).

Both Tharp (1927) and Cheydleur (1929) compared the new and old type exams for testing achievement in modern foreign languages at the college level. In both instances, the old type exams consisted of translation of texts. In Cheydleur's study, the new exams were those produced by Columbia Research Bureau, including multiple-choice, true-false, and completion items. He provides much data comparing the exams for some 1700 students, and comments favorably, even enthusiastically, about the new tests, but in the final analysis one is hard pressed to identify specifically what the relationship was between the two types of tests or between either test and final grades. Tharp shows that the new exam's correlation with the old exam approaches unity (for the corrected r) and that both correlate equally highly with grades; however, Tharp's new exam is entirely of the completion type, thereby not meeting the criteria for this review. A third, curious entry in the foreign language area is the "study" by Lemper (1925), who for one French class (N=28) at Kansas State Agricultural College does not provide any correlations between what he refers to as objective and subjective exams, but actually lists the

scores of all students and comments on them. The author, in general, reacts favorably to the objective exam, with a few exceptions, then concludes "The number of cases in this experiment was so few that objection might be raised to drawing conclusions from the data. Consequently, none will be presented in this article." That being the case, one wonders why the article should have ever appeared.

Several reports (Kinder, 1925; Phillips, 1931; Remmers, et al., 1923; Ruch and Charles, 1928) all seem to have collected data which would allow relating scores on free-response (either essay or completion) items to a variety of choice-type items, but limit themselves to reporting reliabilities and/or difficulty levels.

Finally, Meyer (1935), while referring to an "experimental" study of the old and new types of exams, actually provides only student self-reports of how the various exams affect study habits.

Literature Reviews. It seems appropriate to conclude our review of studies from the early years by presenting three reviews of the literature which were published during that era. Each of these reviews covered much more than the question of interest in this paper, vis. whether free-response and choice-type tests measure the same functions, including the now familiar themes of relative reliabilities, amount of content coverage per unit of time, effect of corrections for guessing, student reactions, etc., but we will limit our treatment of the reviews to our main concern. The reviews cover many of the same articles already discussed in this paper, plus certain studies not generally accessible today. By way of preview, we might note that the three reviews show a remarkable degree of unanimity in their main conclusion, i.e. that the free-response and choice-type tests do indeed measure the same traits, abilities, or functions.

The first summary was published by Ruch (1929), whom we have already noted was himself one of the leading researchers of this question. After detailed

review of five "Studies of Comparative Validity," Ruch (1929, p. 290) provides the following summary:

> 1. Where old- and new type tests are compared, the new type are at least as valid as the traditional examination.
>
> 2. There is no reason to believe that the newer objective tests are impotent for the measurement of reasoning and thought in contrast with memory for facts.
>
> 3. If recall tests are held to be valid (and there is no evidence to the contrary), recognition tests measure roughly the same abilities or functions.
>
> 4. When validities are measured against school marks as a criterion, the correlations are lower than where long objective tests are used as the criterion of validity. Such a finding, however, is in line with the expectancy, since school marks are very unreliable and hence will not support high correlations.
>
> 5. Instructions against guessing seem to give more valid results than where pupils are directed to guess.
>
> 6. When validity coefficients are corrected for attenuation (errors due to unreliability of measurement), the resulting values are high, showing that true-false, multiple-choice, and recall tests measure roughly the same abilities.

The second review was provided by Kinney and Eurich (1932). They summarized 13 studies, noting that (p. 541): "Considerable ingenuity has been exercised in devising criteria for determining the validity of tests." They conclude, after identifying some minor differences in the validities of different types of new tests (pp. 541-542): "It may be stated with considerable assurance that the new type test has been shown to be at least as valid as the essay examination."

Finally, Lee and Symonds (1933) conclude their extensive review on "Comparative Validities" thus (p. 25): "In general these conclusions do not overlap with Ruch's [1929] except in so far as the earlier studies show that the new type are at least as valid as the essay tests and that when the correlations between the essay and objective tests are corrected for errors of measurement, they measure approximately the same abilities." Lee and Symonds' quibble with some of Ruch's conclusions relates to some rather minor differences in comparative validities of various forms of new type tests, e.g.

true-false vs. multiple-choice.

The "Can't Let Go" Syndrome. We referred in the introduction to this review to an overwhelming inclination to believe that the free-response mode of testing was in some sense naturally superior to choice-type testing and that the latter probably missed some special abilities which the former mysteriously captured. Just how overwhelming this inclination seems to be is indicated in part by the fact that a number of the most prominent early advocates of the "new" tests, after rather thoroughly demolishing the supposition that the new and old tests measure different things--and having shown that the new tests are both more reliable and easier to score--these same authors recommend continued use of essay or other free-response measures, albeit in combination with the new tests. Thus, for example, we find Ruch (1929, p. 285) commenting ". . . that recall and recognition tests undoubtedly measure somewhat different abilities;" and Wood (1923, p. 199) noting "One is justified in concluding that a combination of the New and Old methods affords for the present the safest starting point." Nor should it be thought that rabid attacks on supposed evils of objective tests are a phenomenon of just recent vintage; few criticisms of late have been any more strident or very different in character than that of Cason (1931).

MORE RECENT INVESTIGATIONS

There is a remarkable hiatus in the literature on the relationship between free-response and choice-type tests from the late 1930's to the mid 1950's. It may be that researchers in this interim period considered the basic questions to have been satisfactorily answered and, hence, not in need of further study. The hiatus to which we refer, however, consists of more than just an absence of studies: It extends to a lack of references in most later studies to the prodigous literature of the earlier era, almost as if the question of interest had never been raised before. So, for example, we find

statements in the more recent literature such as ". . . it is still uncertain to what extent essay and objective tests over the same instructional unit measure the same or different abilities and thus reflect parallel results after factors of unreliability are removed." It is not at all unusual for recent publications to make no reference whatsoever to studies of the earlier years. Nonetheless (although we hate to destroy the suspense), the more recent investigations largely confirm conclusions from the earlier studies. But (to reestablish some suspense) there are a number of notable nuances in the more recent studies, some of which suggest avenues for further work.

Direct Correlation Studies. As in the earlier years, the most frequent attack in recent years on the question of the equivalence of free-response and choice-type measures has involved the direct correlation of one type of test with the other. Cook (1955) correlated free-response and several choice-type tests for 303 university freshmen. All disattenuated r's exceeded .95. (We have generally avoided referencing dissertations the results of which are not published elsewhere, but an exception has been made for the Cook study because it has been frequently cited in other literature.) Harke, Herron, and Lefter (1972) compared performance of 170 students on a multiple-choice test and the "universally-accepted written solution test" involving work with physics problems. A disattenuated r of .92 was obtained, leading the authors to conclude that the multiple-choice format was an adequate substitute for the written solution test. In an article eerily reminiscent in tone to that of Magill (1934), Colgan (1977) reports "strong positive correlation" between multiple-choice tests and "ultimate marks," these latter being based mostly on conventional, open-ended math problems but influenced to some extent by some other factors; but, based on examination of deviant cases in bivariate plots, the author then remonstrates with the multiple-choice procedure for not being in total agreement with the "ultimate marks." The author seems unaware of the notion that the correlation is affected by lack of perfect reliability in the

two measures. Presentation of the data is such that one cannot disattenuate the correlations given in the article, hence making it difficult to draw definitive conclusions.

Bracht and Hopkins (1970) tested 279 students in an introductory educational psychology course with multiple-choice items (mostly taken from the little item-banks accompanying textbooks) and essay tests "designed to allow for [their] unique measurement characteristics, i.e. the assessing of 'higher cognitive abilities,' i.e. the ability to apply, analyze, conceive, design, and integrate concepts and segments of subject matter." The authors report (p. 362-363) that ". . . the disattenuated correlations between essay and objective tests . . . all exceeded 1.0." Horn (1966) reached much the same conclusion based on analysis of essay and objective tests for a measurement course. Heim and Watts (1967), while being mainly interested in the now-hackneyed subject of the relative difficulty of open-ended and multiple-choice versions of items with identical stems, did report correlations of .91 and .86 between such alternate versions of a vocabulary test for British naval recruits. The latter r's are uncorrected and it is not difficult to suppose that correcting them for attenuation would yield r's approaching unity. Heim and Watts also reported approximately equivalent correlations beteween the two versions of their test and an intelligence test.

Hurlburt (1954) also used identically stemmed vocabulary items as a basis for studying the relationship between recall and recognition modes of responding. Subjects were 192 grade 9 and 210 grade 11 students. Actually, the two test modes (recall and recognition) each had three subtests: one covering nouns, one verbs, and one adjectives. The author reports correlations, corrected for attenuation, of .70 and .72 between the total recall and the total recognition scores, for grades 9 and 11, respectively. These r's are exceptionally low in comparison to those reported in other studies. More strangely, median correlations, corrected for attenuation,

between subtests (nouns, verbs, adjectives) within type of test are .59 and .69 for recall and recognition forms, respectively. Now one does not usually think of knowledge of nouns, verbs, and adjectives as separate abilities. In addition, average scores given for the subtests do not sum to the total scores for the tests and a number of tables are obviously labelled incorrectly. Altogether one is left feeling uneasy, at best, about the Hurlburt findings.

Davis and Fifer (1959) constructed 45-item free-response and multiple-choice measures of arithmetic reasoning ability, making "[E]very effort . . . to maximize the reasoning and to minimize the computation needed to obtain the solution to each problem" (p. 160). The major focus of attention in the study was the effect of certain item-weighting procedures; however, data on the relationship between the two testing modes are presented incidentally. For a group of 251 pupils in grades 8 and 9 of a school for gifted girls, the correlation between free-response and multiple-choice measures was .69. The alternate-form reliability of the multiple-choice test, calculated for a different group of individuals, but one with means and standard deviations similar to those for the gifted girls, was .68. A reliability figure is not given for the free-response measure, but if its reliability is assumed to be about the same as for the multiple-choice version of the test (if anything, perhaps a generous assumption), then the disattenuated r between the free-response and multiple-choice measures is approximately 1.00. The free-response measure appeared to correlate slightly but nonsignificantly higher than the multiple-choice measure with teacher ratings of pupils' abilities to solve arithmetic reasoning problems. Of course, if the disattenuated r between the two types of tests really is 1.00, then their respective correlations with a third variable could differ only as a result of unreliable variance.

We should mention the extensive but still only partially reported research on the American College Testing Program's "College Outcome Measures Project

(COMP)" work, which we shall attribute to Forrest and Steele (1980). The COMP measures are designed to provide very realistic, broadly applicable tests of the general education component of college degree programs, defined operationally in the areas of Functioning in Social Institutions, Using Science, Using the Arts, Communicating (writing and speaking), Solving Problems, and Clarifying Values. The tests have evolved in three forms: a measurement battery which is entirely free-response, an objective test which is entirely multiple-choice, and a composite exam which is partly free-response and partly multiple-choice; all three forms use the same or comparable stimulus materials. The interrelationships among all these measures have yet to be fully reported (Forrest and Steele, 1981), but inspection of certain data in a draft report (Forrest and Steele, 1980) seems to support the authors' contention that the multiple-choice measures are reasonable substitutes for the free-response measures in all areas except writing, speaking, and perhaps some aspects of problem solving. More detailed analysis of the rich data base provided by COMP for the questions of interest in this review is anxiously awaited.

Vernon (1962) investigated a wide range of variables related to test formats and content, including the relation between multiple-choice and "creative (own-word) responses." (We have, as indicated in the introduction to this paper, generally avoided consideration of studies related to reading test formats, e.g. cloze vs. modified cloze techniques. However, the Vernon study is included because it seems very different from other "reading test" studies both in terms of the types of tests used and in data analysis.) Vernon (p. 274) hypothesized, among other things, that "Tests responded to in own words will correlate more highly among themselves, also multiple-choice tests among themselves, than own word with multiple-choice," even making a rather impassioned plea in defense of ". . . the essay examination for eliciting . . . higher educational qualities . . ." (p. 169). Vernon's

analyses are far-ranging, often exploratory in nature, thus making it difficult to present simple summaries of results. In one place (p. 279), he notes, with evident disappointment, that "A comparison of all the multiple-choice and creative-response correlations does not appear to bear out Hypothesis V [the one stated above] . . . . [and] A further examination was made of the residual correlations after removing content factors by factor analysis . . . If any variance is attributable to this difference in item-form, it can hardly amount to more than 1 or 2 per cent." The correlations referred to show for two different samples of college students (one British, one American) average within-multiple-choice r's of .60 and .43, within-creative-response r's of .48 and .49, and across-mode (multiple-choice vs. creative response) r's of .56 and .46, all of which would yield disattenuated across-mode r's of approximately 1.00. Finally, Vernon concludes (p. 285) "The writing of responses to vocabulary and reading questions by students in their own words (creative type) did not, as had been hypothesized, involve a different ability from the objective or multiple-choice type of response."

The Vernon study just reviewed employed factor analytic techniques, among others, in an abortive attempt to identify test format (free-response vs. multiple-choice) factors. Traub and Fisher (1977) relied primarily on factor analytic methodology in a similar attempt. They used two sets of mathematical reasoning and two sets of verbal comprehension items each cast in three formats--constructed response, standard multiple-choice, and "Coombs multiple-choice" (strike out incorrect responses, as many as you wish). Marker tests for following directions, recall memory, recognition memory, and risk-taking were also employed. The authors concluded that the three types of measures were equivalent for the mathematical reasoning items, but not entirely so for the verbal items, for which a "weak" format factor emerged. Some suspicion, however, must be entertained about this entire study since

examinees--junior high students--received six sets of exactly the same items but in three different formats repeatedly over several weeks, a procedure which the authors acknowledge caused some problems "in sustaining student motivation." (Indeed!) The authors note further that their conclusions must be qualified due to lack of parallelism in measures which were intended to be and should have been parallel for the entire factor analytic line of reasoning to work.

Stake and Sjogren (1964) also applied factor analytic methodology to 39 achievement scores per student for 100 college students in an attempt to identify item format factors. Items included essay, performance, matching, multiple-choice, and completion, used with a variety of topics and instructional modes, the latter being the matter of principal interest in the study. The authors hypothesized that item-type factors would be identified. The hypothesis was not confirmed. "No major factors were interpreted as having been determined by item types. It was concluded that if there were any item-type bias in these data at all, it could be considered negligible" (p. 33).

Ward (1981) provides still another example of the use of factor analysis in an attempt to identify an item format factor. He employed four response types--conventional multiple-choice, single answer free-response, multiple answer free-response, and key list (produce an answer, then find it among a long list of alternates)--across three stimulus types--analogies, sentence completion, and antonyms--in the general content area of verbal knowledge and verbal reasoning. Subjects were 315 paid volunteers, all university students and apparently above average in ability even for university students. The author concluded that the various ". . . item types appear to measure essentially the same abilities regardless of the format in which the test is administered."

Quite recently, attempts have been made to push the applicability of choice-type measures to seemingly absurd limits, i.e. within contexts tailor-made for the measurement of creative or divergent production abilities in which a free mode of responding would seem to be essential. Ward, Fredericksen, and Carlson (1980), also summarized in a popular form by Fredericksen, Ward, and Carlson (1980), derived six scores from a test called Formulating Hypotheses. Generally, examinees were asked to formulate hypotheses that might account for given results of a study; responses are scored for quality, sheer number, and unusualness, with some variations in the exact definition of these basic themes. Although the tasks and types of scores were clearly designed with a free-mode of responding in mind, machine scorable versions of responses were developed for each score by listing nine alternative hypotheses and having examinees choose among these. Free responses and machine scorable versions of the formulating hypotheses test, along with a battery of personality and cognitive tests, were administered to 174 paid volunteer college senior psychology majors intending to pursue graduate work; GRE scores were available for some of these students. Analysis of data depended primarily on factor analytic methodology. The authors conclude that the two types of tests--free-response and machine-scorable--are approximately equivalent with respect to quality scores, but not equivalent with respect to number and unusualness of ideas. Given the method of reporting results in the study, it is difficult to describe exactly how different the two types of tests are in these latter scores. It might be noted that there was a remarkably large amount of specific variance in the various scores derived from the Formulating Hypotheses test, both in free-response and machine-scorable formats. For example, while the median correlation between corresponding scores in the two formats was only .19, the median intercorrelations among subtests within format were only .16 and .14 for machine-scorable and free-response versions, respectively. Nonetheless,

the study provides an interesting set of exercises and array of data; it is hoped that further explorations along these lines will be forthcoming.

Criterion Correlation Studies. Cowles and Hubbard (1952) analyzed the relationship between student standings, as rated by department faculty, and performance on essay and objective tests of knowledge in internal medicine and pharmacology. The study was conducted for the National Board of Medical Examiners and involved 636 students in internal medicine and 546 students in pharmacology. The objective test correlated more highly with department ratings in both areas (in medicine, r's = .37 and .21, respectively; in pharmacology, r's = .49 and .18, respectively). Thompson (1965) correlated grades in college physics courses with scores on (a) the one-hour objective section and (b) the two-hour essay section of the CEEB Advanced Placement Test in Physics for 226 students from 1962 and 222 students from 1963. For several different methods of summarizing the data, predictive validities of the objective and essay sections were nearly equal (r's generally about .30) and, interestingly, the combination of the two sections yielded little improvement in predictive validity over either section alone. Unfortunately, in this study we have no specific information on how grades were awarded in courses; to collect such information, of course, would have been almost impossible since the students were involved in nearly a hundred different courses at 18 different institutions.

Kruglak (1965) gives a perplexing set of results for two sections of each of two university physics courses in which short-answer essay and multiple-choice tests were used. He reports higher reliabilities for the essay test than for the multiple-choice test--a very odd finding--with reliabilities of some multiple-choice tests as low as .36, which prompts one to wonder about the care with which those tests were developed. Correlations between the essay and multiple-choice tests are described as "not very high" or "moderate"; while ranges are given for these r's, for various courses and

sections, the presentation of data is such that one cannot disattenuate the r's. (The author seems unfamiliar with the attenuation problem.) However, Kruglak also reports that the multiple-choice tests correlated more highly than the essay tests with "total achievement in the course." Again, these correlations are uncorrected for differences in reliabilities of the respective measures. Altogether, it seems difficult to draw any firm conclusions from this study.

The Difficulty Issue Revisited. Many of the more recent studies continue to investigate the relative difficulty of free-response and choice-type tests. Some of these studies also report data on the correlation between the two types of measures and have, therefore, been considered above. However, other studies in this category, while apparently having collected data which would allow determination of correlations do not report these, concentrating exclusively on the difficulty issue. Among such studies are those of Frederickson and Sater (1953), Farley (1963), McCloskey and Holland (1976), and Carroll and Carroll (1977).

The Effect of Testing Mode. A number of earlier studies (e.g. Meyer, 1935) investigated the equivalence of free-response and choice-type measures in terms of the effect that use of these tests have on students' methods of study, but limited the data collected to student self-reports. More recently, several studies have attacked this question experimentally.

Sax and Collet (1968) compared the effects of multiple-choice and recall test on achievement for students in two Tests and Measurements classes. Subjects were given three recall tests or three multiple-choice tests and were told to expect the same format for the final exam. Actually, for the final exam, students were randomly assigned to either the recall or multiple-choice formats. The authors' hypothesis that students who had received multiple-choice exams throughout the semester would obtain higher scores on both the multiple-choice and recall final exams was confirmed; hence, use of

multiple-choice tests, according to the authors, serve as an effective motivator for study in comparison to use of the recall test format.

Kumar and Rabinsky (1979) had sixty ninth-grade students read a passage for a test to be taken the next day. One-third of the students were told to expect a recall test, one-third to expect a multiple-choice test, and one-third were left in an ambiguous situation, being told simply that they would be tested for retention of the material. Then, one-half of each group received a multiple-choice test while the other half of each group received a recall test. There was no significant effect due to the set to receive one or the other kind of test.

Gay (1980) attempted to show that repeated use of short answer tests led to better learning of material (in an introductory educational research course) than did repeated use of multiple-choice tests. The hypothesis was confirmed only when a short answer test was used as the criterion, but not when a multiple-choice test served as the criterion. Data were based on only 14 cases in each group, from one class.

Sharma (1970) claimed to show differences in the "effectiveness" of essay, short answer, and adjective test questions for high, middle, and low achievers. However, the definition of "effectiveness" and the description of methodology in general are so obscure that drawing any conclusions from the study is impossible.

The recent investigations of the effect on test type of students' study methods and habits provide an interesting new twist to the question of the equivalence of free-response and choice-type measures. Technically, this issue is not a psychometric one, but it is an important educational one. To date, studies of the issue suggest that the mode of testing does <u>not</u> make a difference in terms of students' study habits and methods. However, currently available studies have been quite limited in sample sizes, subject matter, etc. There seems to be room for considerable expansion in this area of study.

We conclude this section with a study which is difficult to classify in the categories we have been using but which may point to an important new area of investigation. Longstreth (1978) administered essay, multiple-choice, and true-false tests covering content material in a psychology course. The correlation between essay and multiple-choice tests reached the usual level (.99, disattenuated) although r's involving the true-false test were lower. More interestingly, when results were analyzed separately by race (White, Black, Asian) there were significant race X test format interactions. The study is probably too limited in sample size and diversity to warrant detailed consideration here, but it is noteworthy as the only attempt available (with the possible exception of the Peters and Martz (1931) study) which looks at possibly differentated effects of test formats for various subgroups of examinees.

GENERALIZATIONS

This review seems to provide the basis for the following generalizations.

1. In most instances, free-response and choice-type measures are found to be equivalent or nearly equivalent, as defined by their intercorrelation, within the limits of their respective reliabilities. Further, the choice-type measure is nearly always more reliable than the free-response measure and is considerably easier to score. This generalization has been developed across a wide variety of subject matter areas; ways of defining free-response formats (essay, completion, short answer, translations, open-ended problems, etc.) and choice-type formats (a variety of multiple-choice, true-false, matching items, with and without assorted weights, directions, and other variations); using diverse methodologies; across a wide span of time; and, for those who worry about Rosenthal effects, by investigators of many different predispositions.

2. To the extent that free-response and choice-type tests diverge in what they measure and there is some outside criterion by which to judge which of the two is the better measure, the choice-type measure, more frequently than not, proves to be more valid, sometimes as a function of its greater reliability but often even after allowance is made for the lesser reliability of the free-response measure.

3. We have deliberately excluded from our review the communication skills of reading, writing, and speaking and hence do not mean to imply (nor deny) that the latter generalizations apply to these areas. It may be that the generalizations do not apply to certain situations deliberately established to provide free rein to divergent thinking ability where there are no right or even preferred answers, although the evidence on this point seems mixed and more investigation is needed to resolve the issue.

4. The overwhelming majority of studies to date have been based on college students, hence those of (a) above average ability, (b) beyond the years of rapid cognitive development, and (c) from predominantly middle-class, White, Western cultural background (and, perhaps having some other peculiar characteristics, too). With one exception, studies done with other populations do not tend to depart in their conclusions from studies done with college students. However, these other studies are sufficiently few in number to prompt the suggestion that studying the question of interest here with diverse groups is an important area for further study.

5. Evidence collected to date suggests that there are not undesirable side effects, e.g. in terms of students' study habits, resulting from use of choice-type tests.

Bibliography

Anastasi, A. Psychological Testing. (4th Ed.) New York: Macmillan, 1976.

Andrew, M. and Bird, C. A comparison of two new-type questions: recall and recognition. Journal of Educational Psychology, 1938, 29, 175-193.

Arnold, H. L. Analysis of discrepancies between true-false and simple recall examinations. Journal of Educational Psychology, 1927, 18, 414-420.

Baker, F. Automation of test scoring, reporting, and analysis. In Thorndike, R. L. (Ed.) Educational Measurement. (2nd Ed.) Washington, D. C.: American Council on Education, 1971, pp. 202-234.

Bayles, E. E. and Bedell, R. C. A study of comparative validity as shown by a group of objective tests. Journal of Educational Research, 1931, 23, 8-16.

Bird, C. and Andrew, D. M. The comparative validity of new type questions. Journal of Educational Psychology, 1937, 241-258.

Bracht, G. H. and Hopkins, K. D. Comparative validities of essay and objective tests. Research Papers, No. 20. University of Colorado, Laboratory of Educational Research, 1968.

Bracht, G. H. and Hopkins, K. D. The commonality of essay and objective tests of academic achievement. Educational and Psychological Measurement, 1970, 30, 359-364.

Carroll, J. L. and Carroll, J. A. A comparison of the WISC information and arithmetic subtests with a multiple choice procedure using kindergarten, first and second grade children. Psychology in the Schools, 1977, 14, 416-418.

Carter, H. D. and Crone, A. P. The reliability of new-type or objective tests in a normal classroom situation. Journal of Applied Psychology, 1940, 24, 353-368.

Cheydleur, F. The relative reliability of the new and old type modern language examinations. French Review, 1929, 2, 530-550.

Colgan, L. H. Reliability of mathematics multi-choice tests. International Journal of Mathematics Education and Science Technology, 1977, 8, 237-244.

Cook. D. L. An investigation of three aspects of free-response and choice-type tests at the college level. Dis. Abs. 1955, 15, 1351. Pub. No. 12,886.

Corey, S. M. The correlation between new type and essay examination scores, and the relationship between them and intelligence as measured by army alpha. School and Society, 1930, 32, 849-850.

Cowles, J. T. and Hubbard, J. P. A comparative study of essay and objective examinations for medical students. The Journal of Medical Education, 1952, 27(2), 14-17.

Crawford, C. C. and Raynaldo, D. A. Some experimental comparisons of true-false tests and traditional examinations. School Review, 1925, 33, 698-706.

Davis, F. B. and Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 1959, 19, 159-170.

DuBois, P. H. A History of Psychological Testing. Boston: Allyn Bacon, 1970.

Ebel, R. Measuring Educational Achievement. (2nd Ed.) Englewood Cliffs, NJ: Prentice Hall, 1972.

Eurich, A. Four types of examinations compared and evaluated. Journal of Educational Psychology, 1931, 26, 268-278.

Farley, F. H. Further data on multiple-choice versus open-ended estimates of vocabulary. British Journal of Social and Clinical Psychology, 1963, 8, 67-68.

Forrest, A. and Steele, J. Personal communication, 1981.

Forrest, A. and Steele, J. College Outcomes Measures Project: Summary Report of Research & Development 1976-1980. Iowa City: American College Testing Program, 1980.

Forsyth, R. A., and Feldt, L. S. Some theoretical and empirical results related to McNemar's test that the population correlation coefficient corrected for attenuation equals 1.0. American Educational Research Journal, 1970, 7, 197-207.

Fredericksen, N. and Satter, G. A. The construction and validation of an arithmetic computation test. Educational and Psychological Measurement, 1953, 13, 209-227.

Fredericksen, N. Ward, W. C. and Carlson, S. B. Can multiple-choice tests measure creativity? Findings: A Periodical of ETS Research in Postsecondary Education, 1980, VI(1), 1-4.

Gates, A. I. The true false test as a measure of achievement in college courses. Journal of Educational Psychology, 1921, 12, 276-287.

Gay, L. R. The comparative effects of multiple-choice versus short-answer tests on retention. Journal of Educational Measurement, 1980, 17, 45-50.

Guiler, W. S. Validation of methods of testing spelling. Journal of Educational Research, 1929, 20, 181-189.

Guilford, J. P. Psychometric Methods. (2nd Ed.) New York: McGraw-Hill, 1954.

Guilliksen, H. Theory of Mental Tests. New York: Wiley, 1950.

Harke, D. Comparison of a randomized multiple-choice format with a written one-hour physics problem test. Science Education, 1972, 56, 563-565.

Heim, A. W. and Watts, K. P. An experiment on multiple-choice versus open-ended answering in a vocabulary test. British Journal of Educational Psychology, 1967, 37, 339-346.

Horn, J. L. Some characteristics of classroom examinations. Journal of Educational Measurement, 1966, 3, 293-295.

Hurd, A. W. Comparison of short answer and multiple-choice tests covering identical subject content. Journal of Educational Research, 1932, 26, 28-30.

Hurlburt, D. The relative value of recall and recognition techniques for measuring precise knowledge of word meaning - nouns, verbs, adjectives. Journal of Educational Research, 1954, 47, 561-576.

Kinder, J. S. Supplementing our examinations. Education, 1925, 45, 557-566.

Kinney, L. S. and Eurich, A. C. A summary of investigations comparing different types of tests. School and Society, 1932, 36, 540-544.

Knight, F. B. Data on the true-false test as a device for college examinations. Journal of Educational Psychology, 1922, 13, 75-80.

Kruglak, H. Experimental study of multiple-choice and essay tests. American Journal of Physics, 1965, 33, 1036-1041.

Kumar, V. K. and Rabinsky, L. Test mode, test instructions, and retention. Contemporary Educational Psychology, 1979, 4, 211-218.

Laird, D. A. A comparison of the essay and objective type examination. Journal of Educational Psychology, 1923, 14, 123-124.

Lee, J. M. and Symonds, P. M. New-type or objective tests: a summary of recent investigations. Journal of Educational Psychology, 1933, 24, 21-38.

Lemper, L. Objective versus subjective tests in modern languages. Modern Language Journal, 1925 10, 175-177.

Longstreth, L. Level I - Level II abilities as they affect performance of 3 races in the college classroom. Journal of Educational Psychology, 1978, 70, 289-297.

Lord, F. M. and Novick, M. Statistical Theories of Mental Test Scores. Reading, Mass: Addison-Wesley, 1968.

Lord, F. M. A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. Psychometrika, 1957, 22, 207-220.

Lord, F. M. Testing if two measuring procedures measure the same dimension. Psychological Bulletin, 1973, 79, 71-72.

Magill, W. H. The influence of the form of item on the validity of achievement tests. Journal of Educational Psychology, 1934, 25, 21-28.

McCall, W. T. How to Measure in Education. New York: Macmillan, 1922.

McCloskey, D. I. and Holland, A. B. A comparison of student performances in essay-type and multiple choice questions. Medical education, 1976, 10, 382-385.

McNemar, Q. Attenuation and interaction. Psychometrika, 1958, 23, 259-266.

Meyer, G. An experimental study of the old and new types of examination: II methods of study. Journal of Educational Psychology, 1935, 26, 30-40.

Odell, C. W. Traditional Examinations and New Type Tests. New York: Century, 1928.

Paterson, D. G. Do new and old type examinations measure different mental functions? School and Society, 1926, 24, 246-248.

Paterson, D. G. and Langlie, T. A. Empirical data on the scoring of true-false tests. Journal of Applied Psychology, 1925, 9, 339-348.

Peters, C. C. and Martz, H. B. A study of the validity of various types of examinations. School and Society, 1931, 33, 336-338.

Phillips, D. P. Comparison of the two-response and dictated recall types of spelling tests. Journal of Educational Research, 1931, 23, 17-24.

Remmers, H. H., Marschat, L. E., Brown, A. and Chapman, I. An experimental study of the relative difficulty of true-false, multiple-choice, and incomplete-sentence types of examination questions. Journal of Educational Psychology, 1923, 14, 367-372.

Robertson, G. Development of the first group mental ability test. In Bracht, G. H., Hopkins, K. D. and Stanley, J. C. Perspectives in Educational and Psychological Measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972, pp. 183-190.

Ruch, G. M. The Improvement of the Written Examination. Chicago: Scott, Foresman, 1924.

Ruch, G. M. The Objective or New-Type Examination: An Introduction to Educational Measurement. Chicago: Scott, Foresman, 1929.

Ruch, G. M. and Charles, J. W. A comparison of five types of objective tests in elementary psychology. Journal of Applied Psychology, 1928, 12, 398-403.

Ruch, G. M. and Rice, G. A. Specimen Objective Examination (A collection of Examinations Awarded Prizes in a National Contest in the Construction of Objective or New-Type Examinations, 1927-28) Chicago: Scott, Foresman, 1930.

Ruch, G. M. and Stoddard, G. D. Comparative reliabilities of five types of objective examinations. Journal of Educational Psychology, 1925, 16, 89-103.

Ruch, G. M. and Stoddard, G. D. Tests and Measurements in High School Instruction. Yonkers, NY: World Book, 1927.

Sax, G. and Collet, L. An empirical comparison of the effects of recall and multiple-choice tests on student achievement. Journal of Educational Measurement, 1968, 5, 169-173.

School and Society. Experimenting with the new type of examination at Columbia. School and Society, 1922, 15, 141-142.

Sharma, V. P. Efficacy of evaluation procedures in relation to pupils' scholastic attainment. Indian Psychological Review, 1970, 6, 107-109.

Shulson, V. and Crawford, C. C. Experimental comparison of true-false and completion tests. Journal of Educational Psychology, 1978, 19, 580-583.

Stake, R. E. and Sjogren, D. D. Activity level and learning effectiveness. NDEA Title VII, Project Number 753. University of Nebraska, 1964.

Stanley, J. C. Reliability. In Thorndike, R. L. (Ed.) Educational Measurement. (2nd Ed.) Washington, D. C.: American Council on Education, 1971, pp. 356-442.

Tharp, J. B. The new examination versus the old in foreign languages. School and Society, 1927, 26, 691-694.

Thompson, R. E. A study of the comparative validities of the essay and objective sections of the C.E.E.B. Advanced Placement Examination in Physics (TDR-65-4). Princeton: Educational Testing Service, 1965.

Toops, H. A. Trade Tests in Education. (Teachers College Contributions to Education No. 115.) New York: Teachers College, Columbia University, 1921.

Traub, R. E. and Fisher, C. W. On the equivalence of constructed response and multiple-choice tests. Applied Psychological Measurement, 1977, 1, 355-369.

Vernon, P. E. The determinants of reading comprehension. Educational and Psychological Measurement, 1962, 22, 269-286.

Ward, W. C. Measurement of aptitude for divergent verbal production. (Multilith draft.) Princeton, NJ: Educational Testing Service, 1981.

Ward, W. C., Fredericksen, N., and Carlson, S.B. Constant validity of free-response and machine-scorable forms of a test. Journal of Educational Measurement, 1980, 17(1), 11-29.

Weidemann, C. C. and Newens, L. F. Does the "compare-and-contrast" essay test measure the same mental functions as the true-false test? Journal of General Psychology, 1933, 9, 430-449.

Wood, B. D. Measurement in Higher Education. Yonkers, NY: World Book, 1923.

Wood, B. D. New York Experiments with New-Type Modern Language Tests. [Publications of the American and Canadian Committees on Modern Languages] NY: Macmillan, 1927.