Perspective on Educational Measurement

Harold Gulliksen
Emeritus, Princeton University, and Educational Testing Service

An important but usually neglected aspect of the training of teachers is instruction in the art of writing good classroom tests. Such training should emphasize various forms of objective items (e.g., multiple-choice, master list, matching, greater-less-same, best-worst answer, and matrix format). The proper formulation and accurate grading of essay items should be included, as should the use of various types of free-answer items (e.g., the brief answer, interlinear, and "fill in the blanks in the following paragraph" forms). For courses involving laboratory work, such as science, machine shop, and home economics, performance and identification tests based on the laboratory work should be used.

A second point is that organizations developing aptitude tests for nonacademic areas, such as police work, fire fighting, and licensing tests, should emphasize the use by the client of a valid, reliable, and unbiased criterion. Organizations developing academic aptitude tests should also (1) be alert to the accuracy of criterion measures, grades, rank in class, and so forth; (2) call teachers' attention to defects in grading; and (3) help guide teachers and schools in improving these procedures. In recent decades, there have been few instances in which a testing organization has apprised teachers of the fact that their criteria—among others, grades on tests and student papers—are often quite unreliable based on characteristics such as work habits and attitude in class, and could be improved by using better tests to evaluate student performance. Characteristics of the group used for determining validity are also critical.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 2, June 1986, pp. 109–132

0146-6216/86/020109-24$2.45

It is the purpose of this paper to make two recommendations related to testing: (1) that teacher training institutions should emphasize and expand the teaching of test construction for classroom use, and (2) that makers of academic tests, both standardized and individualized, should develop tests on the basis of valid, reliable, and unbiased performance criteria, as in occupational and professional testing programs.

Persons concerned with academic learning, testing, and evaluation must bear in mind that since standardized testing might require the frequent administration of a test to thousands of students at a large number of different schools in different localities over a period of time, such a test can measure only the objectives believed to be common to all. Individualized teacher-designed classroom tests, on the other hand, should be given on a daily, weekly, or monthly basis, and may cover material from a single sector of the course—such as one principle of grammar, or a single chapter or study unit (e.g., an historic period). Such tests should be scored and returned to the students within a week or a single day, or could even be scored in class by the students themselves and used for immediate class discussion.

The failure to distinguish between the requirements of standardized testing and classroom testing seems to be responsible for the lack of improvement—and perhaps even a decline—in the quality of teacher-made classroom tests over the last 40 years.

Traditionally, the nature and methodology of classroom test construction receive short shrift in teacher training institutions and in education texts. Textbooks on item writing composed during the 1960s and 1970s assert that there are three types of items—essay, true-false, and multiple-choice—and then devote a chapter to each type (cf. Ahmann, 1962; Beggs & Lewis, 1975; Bloom, Hastings, & Madaus, 1971; Cronbach & Snow, 1977; Ebel, 1965; Furst, 1958). Texts of the 1930s and 1940s typically described many more types of objective test items, such as master list, matching, greater-less-same, best-worst answer, and matrix format (cf. Adkins, 1974; Adkins, Primoff, McAdoo, Bridges, & Forer, 1947; Burt, 1948; Hawkes, Lindquist, & Mann, 1936; Monroe, DeVoss, & Kelly, 1924; Richardson, Russell, Stalnaker, & Thurstone, 1933; Rinsland, 1937; Ruch, 1929). There are also various forms of free-answer items in addition to the essay, such as the brief answer, comment on the following statements, interlinear, and fill-in-the-blanks; these are all very useful types of items, and are especially suited to classroom testing. Carlson's (1985) handbook, *Creative Classroom Testing: Ten Designs for Assessment and Instruction,* was prepared with the help of a number of teachers and presents a number of objective item types that are not multiple-choice. In laboratory courses, such as science, machine shop, and home economics, performance and identification tests involving equipment, materials, and procedures add to the variety of instruments available to the teacher.

The second recommendation of this paper, that educators emulate the practice of occupational and professional test-makers in emphasizing that the client develop and use valid, reliable, and unbiased criteria for measuring performance objectives, is based on the success of such nonacademic programs in predicting job performance. Test developers for tests involving police work, fire fighting, real estate licensing, etc., usually do emphasize the development and use by the client of valid, reliable, and unbiased criteria that are used to evaluate the content and nature of the resulting tests (cf. Rosenfeld & Thornton, 1978; Thornton, 1979; Thornton & Rosenfeld, 1980).

Organizations charged with developing academic aptitude tests should be alert to the accuracy of criterion measures, and should advise teachers when such measures are found to be deficient. Rarely in recent decades has a testing organization called the attention of teachers to the possibility that their grades (including marks on student papers) may be unreliable, or may be based on aspects of student performance other than subject knowledge, such as class participation, conscientiousness, and effort. Evaluation could be improved considerably by the use of better tests designed to measure the objectives of a particular teacher in a particular classroom.

Let us look at some of the major events in the last fifty years of the history of test construction which have led me to make the two main recommendations offered in this paper.

Test Construction

The University of Chicago Board of Examinations

About 1930, President Robert Maynard Hutchins introduced an examination system at the University of Chicago. The procedures developed by the Chicago Board of Examinations during the 1930s for the first two years of college are also applicable at lower grade levels. The curriculum for the freshman and sophomore years consisted of 5 one-year courses in biological science, physical science, social science, humanities, and English. Passing each of these courses required successful completion of a six-hour exam.

Initially in 1930, Louis L. Thurstone was appointed chief examiner; Marion Richardson was examiner in physical sciences, James Thomas Russell in biological sciences, and John Stalnaker in humanities and English (Russell and Stalnaker were examiners from 1931–1936). I was examiner in social sciences from 1934 to 1940. Others later associated with the examining office were Dael Lee Wolfle (biological sciences), George Frederic Kuder, Dorothy Adkins, and Ben Bloom. In the late 1930s, Ralph Tyler replaced Thurstone as chief examiner. It should be noted that in 1947, after

Dorothy Adkins had gone to work for the Civil Service Commission, she and others prepared a very good book (Adkins et al., 1947) on constructing objective and performance tests.

Test security versus disclosure. One of the first rules Thurstone established was: "The day after an exam is given it goes on sale in the University of Chicago Bookstore." The issue of "test security" versus "disclosure" has now become a legal issue.

In general, for a given instructor's testing of his or her own classes, maintenance of "secure" items (e.g., items used on previous tests where difficulty and correlation with total test score are known, but have not become available to students) is an inappropriate policy. Students should be informed as fully as possible regarding the course requirements, the nature of the tests, and the skills and knowledge they are expected to have gained as a result of taking the course. Making previous exams available to both present and future students is one way to achieve this objective. Also, making previous exams available prevents special advantages from accruing to certain groups (e.g., fraternities, sororities, special tutors, or coaching schools) that will, from time to time, be able to obtain access to test material that the instructor is attempting to keep secure. When instructors do not depend on secure items for equating of tests or grades from year to year, then it is necessary to depend on instructor judgment regarding similar difficulty of parallel items.

The problem of equating grades and tests that show improvement (increased scores) from one year to another also taxes the instructor's judgment, as in cases where certain material is missed by many students one year and special attention is paid to ensure that students learn the material in subsequent years. When such improvement (increase) in scores occurs, the instructors must decide the extent to which grades will be increased to reflect this improvement versus raising of standards. Of course, a corresponding decision in the reverse direction is necessary whenever student performance declines.

Exams and item types. At the University of Chicago the examiners and teaching faculty together constructed the examination items, which were then reviewed, revised, and approved by the teaching staff in each course. The written exam consisted of objective items of various types, such as master list, true-false, greater-less-same, best-worst answer, rank order, and tabular (matrix), as illustrated below.

In addition to these easily scored objective item types, various free-answer types were also used, such as short-answer items (two or three to a page), put a word in the blank, correct the bracketed portions of the paragraph, and a one- or two-page essay that was the last item in both the morning and afternoon portions of the exam. The bulk of the exam, however, consisted of objective items. In many instances it was possible to construct a set of items with a mutually exclusive and exhaustive list of all possible alternatives (e.g., increases, decreases, stays the same, not enough information to determine). Thus, it was not necessary to construct plausible distractors for each question.

The exams at Chicago were not composed entirely of written items. Where laboratory work was important, as in the biological and physical sciences, the final exam involved laboratory setups. In the laboratory class we noticed that the instructor spent a fair amount of time moving around the laboratory and pointing out to students various things that were wrong with the apparatus they were using. The students corrected these errors when the instructor pointed them out. Consequently, we set up a number of laboratory experiments; the students were to observe the setup, and either write down what was wrong or indicate that the setup was proper and ready to use. Also, several of the stations would involve the actual performance of an experiment or part of an experiment, such as dissecting a frog's leg for a nerve-muscle preparation. A laboratory assistant was present at each station to give general directions and to grade the performance; the lab assistant was given prior instruction in how to phrase the directions to the student, as well as the points to watch for in grading. Similarly, in addition to the written exams in physics and chemistry, apparatus setups would be criticized and an experiment or two performed as part of the final examination.

These test items proved to be excellent teaching

and learning devices, since they helped make it clear to the students just what they were expected to learn to do in the laboratory. Today computers could be used for simulations of these tasks, with the student making the appropriate keyboard entries to perform the corrections. However, the computer is not necessary for such test items. A regular lab experiment can be set up for each member of the class even when many computers are not available, and the actual equipment may give a more realistic item than the computer would.

At Chicago, the reliability of the total exam was determined by correlating the morning and afternoon scores. The various parts of the exam were also evaluated by correlating the two essays, parallel content pages, etc. Short-answer items, with answers given by the students, were saved, and objective best-worst answer items were constructed for a later exam. Then the free-answer form was given in the morning and the objective version in the afternoon, and the two forms were correlated; this revealed to the faculty the extent of agreement between free-answer and objective versions of the same question. The agreement was very high between the free-answer and the best-worst answer, serving to demonstrate to the faculty that objective items did not damage the evaluative power of the exam. Some of the free-answer items were graded independently by two persons, usually on a five-point scale (A, B, C, D, and E), and the results were plotted and shown to the faculty. Any disagreement between the two readers usually astonished the faculty and helped remove objections to using objective items.

Alternatives to Multiple-Choice Items

Textbooks of the 1930s and 1940s typically describe many types of test questions besides multiple-choice questions, but by the 1960s and 1970s, test developers seem to have settled for three types—essay, true-false, and multiple choice. And they tell us that writing 10 items in an eight-hour day is good productivity. Writing 10 usable multiple-choice items can easily take an eight-hour workday, considering that four plausible false answers must be devised as well as the stem and true answer; but it is difficult to imagine any working teacher, faced with a classroom of lively students, with the leisure to design a routine testing instrument at such a rate. Other types of items, such as master list and matrix format (where all or most of the statements are true), can be written by a person who knows the field at a rate of about 20 items per hour (Hawkes et al., 1936).

A part of teacher training in schools of education should be the writing of items of various types, including free-answer and objective as well as performance and identification items, emphasizing the distinctions and the skills that the teacher is trying to impart. To give teachers the idea that the use of computers and machine-scoring for classroom tests is desirable, or that machine-scorable item types are the best or the only ones to use, will inhibit progress. With over 40 million public school students in the United States, and only about 600,000 or 700,000 computers, the important thing to stress is what teachers can accomplish without computers.

Teacher-made classroom tests should be used frequently, for instance when the class has finished a chapter of a book or a segment of a course, to make clear to the students what they should have learned in a specific part of the course. It might also be a good idea to give such a test before the class studies the material, partly to see how much of it they already know, and partly to give them a clear-cut idea of what they are supposed to be learning. In the case of individualized instruction, where each student proceeds at his or her own pace, the same type of pretest and posttest procedure would be valuable.

Over the last 35 to 40 years, the quality of instruction of teachers to write good exam items for their classes has probably declined rather than improved. William Turnbull, a few years before retiring from the presidency of Educational Testing Service (ETS), agreed with this appraisal and set up a committee under John Helmick to prepare a manual on item writing that could be used by teachers interested in writing better test items. As mentioned above, Sybil Carlson took over the job and prepared such a handbook (Carlson, 1985).

In one very good course and exam in radio given

in high school, the instructor used an interesting procedure. A month or two before the end of the course, he handed out a list of 100 or 200 questions and said, "Your final exam will be 10 of these." All of the students worked hard to learn the answers to the questions, and as a result, they learned what the instructor wanted them to learn in the course. Similarly, giving the students a large number of items and saying, "The final exam will be 50 or 100 of these," would also be a good teaching procedure. The teacher should write the items over a period of time and not accept them from an outside source. For some items, such as translating from or into a foreign language or interpreting graphs, similar items of the same type may be substituted for the study items in the final form of the test.

In discussing exams, teachers frequently say: "I can write a stem and a correct completion. That is no problem. But four plausible false completions are extremely difficult and time-consuming." They should be told to give a free-answer exam first, and then use the students' wrong answers for the distractors in the objective form given later. Some texts say that having the students provide the false completions is impractical, but it was done routinely at the University of Chicago and was found to be a convenient and time-saving practice.

The best-worst answer item type (see Figure 1) is not presented in Carlson (1985) because ETS reviewers felt that a worst answer would be impossibly difficult to write. Only best-answer items were included. Because best-worst answer gives twice as many responses as best-answer, it is very probable that its reliability and validity would be higher. Also, it requires the student to make judgments similar to those the faculty makes in judging the answers to the parallel free-response item. The faculty would very likely resist grading on a two-point scale (A versus BCDE, or AB versus CDE). Grading the items on a three- or five-point scale requires the student to make judgments similar to those made by the faculty in grading items, and

Figure 1
Sample Item: Best-Worst Answer
(Objective Form Constructed From Students' Responses to Free Answer Form)

Below is a question with three answers.
Mark a plus (+) for the best answer.
a minus (-) for the worst answer.
Leave the other answer blank.

Question: What is the relation of hereditary mental traits of a race to the progress of that race in civilization?

Answers

_____ There is hardly any evidence to show that the anatomical characteristics of the races which possess the highest civilization are different from those of races which are on lower levels of culture; but there may be some justification for the belief that hereditary mental traits are to some extent correlated with culture level.

_____ Since there is no good criterion of cultural advancement of a race, and we know little about the relationship between mental and anatomical traits, the most probable view is that there is no necessary connecton among the three variables, hereditary mental traits, anatomical traits, and culture level.

_____ Actually there are observable correlations between hereditary mental traits, anatomical traits, and culture level, and a causal relation exists between these things.

demonstrates more clearly that such objective items require judgments essentially the same as essay items. Faculty grading of a free-answer item constitutes an objective item for the faculty members who grade it. Note that each faculty member grading the item studies the same material, and then responds A, B, C, D, or E.

Teachers should also be told to try various other item types. They should simply begin by asking: "What do I expect the students to differentiate after having my course that they could not differentiate before?" One answer might be, "I expect them to know the difference between the views of John S. Mill, Adam Smith, John Kenneth Galbraith, Karl Marx, Frederick Engels, and Lenin." These names would constitute the *master list*, and it would be followed by a series of statements, every one of which would be easy to write for anyone who knows the field because each would be a true statement. The student's problem is to assign the statement to the person or persons whose view it expresses. Because all the statements are true, the task of devising plausible false statements is obviated.

Of course, if plausible false statements occur to the teacher, there is no reason to reject them. Simply include them in the list and add "none of the above" to the master list of choices. It is necessary to specify in the directions that "Only one name is correct for each statement," or "A statement may express the view of several of these persons or of none of them." Figure 2 is an example of a master list or key list item. Often the matrix format is useful for such items because it provides for multiple answers and for omission of ambiguities. Figure 3 is an example of a matrix format item. The matrix format provides a very convenient method (i.e., crosshatching out an alternative) for removing items that the faculty decides are ambiguous.

Student knowledge of an experiment which had been included in assigned reading can be tested by giving a brief description of the experiment and then listing a number of possible results, as illustrated in Figure 4. Ability to construct a coherent explanatory paragraph can be tested by a rank-order item, as illustrated in Figure 5.

An item to test the students' knowledge of trends and sequences is illustrated in Figure 6. Anyone who knows the field can write 20 items of this type in an hour—not in two 8-hour days. Preparing and grading such items takes far less time than preparing and grading a set of essay items over the same field for a class of 10 or 15 students. For smaller classes, preparing and grading an essay test might take less time.

Another of the item types used in the Chicago program was "Comment on the following statement," followed by a short quotation from a public figure—either a good statement that the student could endorse, or an ill-advised one that the student who had learned the points of view developed in the course would reject, giving justifications. A similar type of item was "Give a brief answer to the following question," presented three to a page with about eight lines allowed for each answer. Illustrations of additional item types are available (see Adkins, 1974; Adkins et al., 1947; Carlson, 1985; Gulliksen, 1985; Hawkes et al., 1936; Richardson et al., 1933).

Test Validity

The College Entrance Examination Board and College Admissions Requirements

When the College Entrance Examination Board (CEEB) was founded in 1900, there was distinct emphasis on the desirability of having the testing program influence the training given in the secondary schools. During the later 1800s, the secondary schools had found it increasingly difficult to prepare students for such schools as Harvard, Yale, Columbia, or Princeton, because the requirements were different for each—specifying certain chapters in given Greek and Latin texts in one institution, and other chapters in the same books or different books in another institution. It was very difficult to teach a secondary school class of students who wished to prepare for different universities, and impossible to teach a given student so that he or she might be admitted to any one of several institutions.

In 1914, when President Butler of Columbia resigned as CEEB Chairman, he said in his parting speech:

> The College Entrance Examination Board has shown that examinations may be so improved,

Figure 2
Sample Item: Master List or Keylist

In the blank space before each of the following statements write the number of the one best term.

1	transference	6	introversion
2	regression	7	extroversion
3	projecton	8	ambivalence
4	manifest content	9	conversion
5	latent content		

4 The subject reported a dream of being lost in a snowstorm, and of being very cold, especially on his hands and feet.

2 Soon after he lost his job, Mr. X, who had spoken excellent English for years, could only speak Polish, which was the language of his childhood.

2 Is characterized by withdrawal tendencies and the adoption of habit patterns long since discarded.

1 During the course of the analysis all the resentment which Ruth had felt for her parents was felt toward the psychoanalyist.

9 A bookkeeper who disliked his job suddenly developed a functional blindness.

7 Trust no Future, howe'er pleasant!
Let the dead Past bury its dead!
Act--act, in the living Present!
Heart within, and God o'erhead!

3 The pot calls the kettle black.

5 An analysis of the dream showed that it referred to a basic conflict between the child and its parents.

6 Here where the world is quiet, here where all trouble seems
Dead winds and spent waves riot, in doubtful dream of dreams.

8 Yet each man kills the thing he loves, by each let this be heard
Some do it with a bitter look, some with a flattering word.

and their effect on secondary school instruction made so stimulating that for a college to maintain separate admission examinations of its own is surely a mark either of weakness, or of perversity, or of mere parochialism, or of the stubborn persistence of educational inertia. (Fuess, 1950, pp. 75–76)

During the late 1930s and early 1940s with the increasing use of an objective Scholastic Aptitude Test (SAT) and the introduction of objective achievement tests in various subject areas, the emphasis on the desirability of centralized examinations guiding or dictating education in secondary schools decreased, because objective exams could cover a much wider range of material than an essay exam. It was stressed that essay examinations based on definite syllabi had been responsible for "a dictated and controlled secondary school curriculum"

Figure 3
Sample Item: Tabular or Matrix

Write the appropriate number or numbers in each space. (Do not put any number in spaces which have been hatched out)

PORTAL OF ENTRY	CAUSAL ORGANISM	AGENTS OR MEANS OF TRANSMISSION	METHOD OF CONTROL
1. skin wound	1. virus	1. mosquito	1. screening homes
2. respiratory tract	2. bacterium	2. flea	2. killing rats
3. urogenital tract	3. protozoa	3. dog	3. vaccination or inoculation
4. intestinal tract	0. none of these	4. man	4. quinine
0. none of these		5. water	5. antitoxin
		6. cow	6. draining swamps
		7. milk	7. pasteurize milk
		0. none of these	8. purifying water
			0. none of these

DISEASE	PORTAL OF ENTRY	CAUSAL ORGANISM	AGENTS OR MEANS OF TRANSMISSION	METHOD OF CONTROL
Malaria				
Smallpox				
Syphilis				
Typhoid				
Rabies				
Diphtheria				
Beri-beri				

(Fuess, 1950, p. 169). It was pointed out that "some examiners had exercised a police function in using outworn definitions of requirements as instruments of institutional control" (Fuess, 1950, p. 115). In contrast, there was now a new emphasis on allowing secondary schools greater freedom in determination of curriculum.

These advantages of objective tests were reinforced when the manpower requirements of World War II made it impossible to assemble enough teachers in the summer of 1942 to grade the essay examinations. Accordingly, objective exams were substituted for the essays in that year, as a temporary wartime measure. By the end of the war, the colleges had found that the results they were getting from the objective tests were serving them as well as those obtained previously from the essay tests. This fact, coupled with the advantages of speed and economy offered by the objective tests and the curricular freedom they offered, led to the adoption of the objective tests on a long-term basis. In the 46th ETS Annual Report, Henry Chauncey pointed out the benefits of the "sudden transition in 1942 from essay exams based on a definite syl-

Figure 4
Sample Item: True-False

In a study of reasoning Maier trained rats in the following way:

(1) On some days each rat was allowed to explore a table top which had food in corner A, a box in corner B, and a low wall across the middle of the table separating corners A and B. The rats could and did climb over this wall.

(2) On other days each rat was placed in the box in corner B and learned to run a maze leading out from the table and back to corner A.

(3) Rats were placed on the table, but prevented from reaching the food in corner A by a transparent screen.

Indicate your knowledge of the results of this experiment by marking each of the following items:

plus (+) if true,
zero (0) if false.

_____ The rats succeeded in forcing their way through the screen and reached the food.

_____ The rats attempted unsuccessfully at first to get through the screen to the food.

_____ The rats gave up trying to reach the food and rested in a corner of the table.

_____ The rats took the maze pathway to the food, but were obviously disturbed, making many errors on the way.

_____ The rats took the maze pathway to the food, running with essentially no errors.

_____ The behavior of the rats was interpreted as indicating a trial-and-error solution of the problem.

_____ The behavior of the rats was interpreted as indicating frustration and regression.

_____ The behavior of the rats was interpreted as indicating an insight solution.

_____ The rats showed a typical form of neurotic behavior.

labus to objective tests which cover so far as possible the common elements in various schools of the country'' (Fuess, 1950, p. 174). Frank H. Bowles, Director of Admissions at Columbia University, an active leader in CEEB affairs, and later President of CEEB, emphasized that ''each college applies its own standards and considers the test results, not by themselves, but along with the school record, principal's report, interview estimate, and all other information available—[this will] enable

Figure 5
Sample Item: Rank Order

Writing a Logical Paragraph

The sentences given are in proper form to make a coherent paragraph, but they are not in the proper sequence. Arrange them logically, placing the appropriate number next to each sentence.

__2__ A suspension bridge is based on the theory that if cables can be strung across an area, a bridge can be hung from them.

__4__ The roadway platform is supported by vertical cables, which in turn are attached to the dipping cables.

__1__ The most practical kind of bridge for use in spanning wide areas is the suspension bridge.

__7__ This will allow for stabilization of the bridge, and will make it more secure.

__3__ The bridge is made of sturdy steel cables which are supported by towers and secured in the ground at either end.

__5__ Girders then support the roadway, in order to prevent the bridge from movement.

__6__ According to most engineers, the main span of a suspension bridge should not be more than 7,000 feet in length.

the schools in formulating their educational program to take advantage of the freedom permitted by the new tests'' (Fuess, 1950, p. 196).

The ''Committee of Revision had broad authority to supervise, review, and coordinate the examination policies of the Board,'' and its functions included ''the analyzing subject by subject of the Board examinations in order to discover new methods of increasing their validity and reliability'' (Fuess, 1950, p. 116). In other words, what the secondary schools were doing in their design of curriculum and teaching was assumed to be correct. It was up to the College Board to alter its tests so that they were ''valid'' in terms of a high correlation with grades in college. This was the general view, with only a few persons giving cautions, such as John M. Stalnaker, who in 1944 said that ''low correlations between test scores and course grades do not necessarily indicate that a change in the tests is desirable. Several other factors must also be considered'' (Fuess, 1950, p. 161).

Validity of Aptitude Tests May Indicate a Faulty Criterion

In evaluating validity coefficients of aptitude tests, the routine procedure is to regard high validity coefficients as indicating that the aptitude test is good and should be used, while low validity coefficients indicate that the aptitude test should be discarded or revised. Only rarely is the question raised, ''Should this validity be high or low?'' There have been a few exceptions to this attitude.

Wesman (1950) emphasized, with several illustrations, that a low validity coefficient may be very useful and enlightening, stimulating a more careful examination of the criterion, and that as a result, the criterion may be revised. The following five illustrations are given by Wesman:

I. A numerical ability test was given to a group of ninth grade boys at the beginning of a school year. . . . At the end of the year . . . the guidance director com-

Figure 6

Sample Item: Greater-Less-Same, or Before-After-Same

The Normal Course of Motor Development (birth to age 3)

The following paired statements refer to milestones in infant motor development that generally are acquired sequentially. If the motor skill described on the left appears before the motor skill described on the right, encircle the letter "B"; if the skill on the left appears after the skill on the right, encircle the letter "A"; if the skills on the right and left appear at approximately the same time, encircle the letter "S".

Skill	When does it appear?	Skill
	A	
head control	B	rolling over
	S	
	A	
sitting up	B	crawling
	S	
	A	
turning from back to stomach	B	turning from stomach to back
	S	

puted the correlation of the students' numerical ability scores with their geometry grades for the year. The coefficient found was about .30. . . . he looked up the students' scores on a statewide examination in geometry, and correlated these scores with the numerical ability test scores. The coefficient in this case was over .60—relatively good prediction. The guidance director used the discrepancy between these two correlation coefficients to initiate discussions with the mathematics teachers as to the bases on which grades were being assigned. The teachers agreed to rate competence and work habits separately; the test is being retained as a selection device by the school. Incidentally, both the school administration and parents are finding math grades more useful than before.

II. A personnel manager in a large industrial firm . . . gave the test, which involved dictation and transcription at high speed, to all stenographers and secretaries already employed by the organization. He also obtained ratings as to the ability of these employees, and . . . correlated [them] with the scores on the proficiency test. To his consternation, the coefficient was quite low. . . .

First he obtained separate coefficients for those called stenographers (who were part of a pool under a stenographic supervisor) and for those called secretaries (each responsible to one or two executives). The correlation of the proficiency test with the stenographic supervisor's ratings was rewardingly good; for secretaries, it was a little worse than it had been originally. The personnel manager discreetly inquired concerning the bases on which the executives rated their secretaries. He was not entirely surprised to find that such factors as assisting in executive decisions, doing personal shopping for the executive, keeping appointments straight, protecting the executive from undesirable visitors, and so on, were [among the factors] affecting the ratings. . . . The secretaries . . . had lost much of their earlier skill through lack of practice. Consequently, they scored comparatively low on the test, and were rated high by their bosses: the validity coefficient suffered. The personnel manager proceeded to install the test as a selection device for newcomers, with confidence that it would predict well where it needed to—at the stenographer level.

III. In a city which has five high schools, a series of aptitude tests was given to all tenth grade students. Verbal reasoning test scores were correlated with physics grades for all students in the five high schools who had taken that course. The correlation coefficient was quite low, between .15 and .20. . . . The research director . . . computed separate validity coefficients for each of the five schools . . . the five coefficients ranged from between .30 and .50. . . . The five schools varied considerably in the quality of their pupils. . . . Each school gave grades to its students according to their performance as compared with their own classmates. Thus, a performance worth A or B in the poorest school was no better than the performance for which a grade of C or D was assigned in the best school. Under the circumstances, it was inevitable that a low validity coefficient would result when the data from all five schools were combined. . . .

IV. A test in American history was used by an eleventh grade teacher at the recommendation of the supervisor of secondary education. The test correlated poorly with the teacher's grades, and the teacher complained to the supervisor that the test was inappropriate for that school. The supervisor . . . sat down with the history teacher, and they analyzed the test items as to whether they were testing for memory of facts or whether they were measuring more complex thought processes. They then rescored the test papers getting one score for "fact" questions and a separate score for "thought" questions. When the teacher's grades were correlated with these new scores, they found that the correlation with the thought part of the test did not improve, while the teacher's grades correlated much better with the scores on fact questions.

As a result of this analysis, the history teacher . . . realized that, in his own . . . examinations, too much stress had been laid on simple memory for facts and too little on ability to use the facts in thinking. . . .

V. An eighth grade shop class took tests of mechanical reasoning and space relations at the beginning of the year. At the end of the first term, the scores on these tests were correlated with the teacher's grades. The coefficients were .26 and .13, respectively. . . . When the second term's grades were in, correlation coefficients were again computed. The mechanical reasoning test correlated .41; the space relations test, .33. . . . They discovered that the first term's work was almost entirely manipulative. . . . During the second term, on the other hand, they were

expected to carry forward more complicated assignments with considerably less supervision.

. . . The prediction of second term grades was a more reasonable demand on the tests.

. . . Regardless of how high or how low a coefficient of correlation may be, these things always demand consideration:

1. How the judgments were arrived at;
2. The nature of the test tasks and their appropriateness in relation to the job or the course; and
3. The peculiarities of the particular group of individuals being studied.

The correlation coefficient has been likened to a three-legged stool: one leg is the predictor (frequently a test), another is the criterion (grades, ratings, etc.), and the third is the population on which the coefficient is obtained (grade level or job family, sex, spread of ability, etc.). He who uses a three-legged stool without ascertaining that all three legs warrant confidence is very likely to be floored. (pp. 20–22)

Lindquist (1961) emphasized that:

. . . unfortunately correlation coefficients . . . have come to be regarded almost solely as "validity coefficients" . . . or as measures of the . . . accuracy of the *predictions* made of college success. . . . The correlation between any so-called predictor and any so-called criterion may and often does reveal just as much about the criterion as about the predictor. . . . It is the purpose of this study to report intercorrelations among the variables for whatever these relationships may reveal about the student, the high school, and the college, as well as about the tests. . . . An ultimate objective, then, is really . . . to identify and do something about the *other* factors . . . such as variations in grading standards and instructional objectives; variations in emotional, motivational, and environmental factors affecting the student while in college; in reliability of grades; in appropriateness of instructional materials, procedures, and goals with reference to individual student needs; and so on. (pp. 7–9)

Numerous reasons for variability are listed, among them: (1) unreliability of predictors and criterion measures; (2) variation in grading standards from instructor to instructor and course to course within and between institutions; (3) differences in learning experiences, personality, and other attributes of students; (4) differences in range of talent, and (5) chance fluctuations.

Validity of Tests in Navy Service Schools

During World War II, Norman Frederiksen and I, as well as others working on achievement test development in Navy service schools from 1942 to 1945, found numerous instances of recruits not learning, or not being graded on, what the Navy instructors thought they were learning and said they should be learning. Introducing a relevant testing program changed this situation radically. Our first experience in evaluating and changing the pattern of validity coefficients was at the gunner's mate school. The initial study found that the validity of the Reading Test was high, and the validity of the Mechanical Knowledge and Mechanical Comprehension Tests low. Checking on the course procedures for testing and grading the students, we found that a long (approximately 1 hour) written test based on the information given in the manual was used. Clearly, the ability to understand and interpret written material was important in determining course grades.

For the practical section of the test, the class of about 50 students was told to relax in chairs or on the floor. Students were individually called to the instructor's table, upon which were a large number of disassembled gun parts. The instructor picked up one, handed it to the student, and said, "What is that called? What is its function?" The same question was asked for a second part, selected at random from those on the desk. The student was then told to return to his seat, and the next one came up and was asked about two other parts. After about an hour and a half of testing, each student had answered questions about two gun parts. This constituted the practical section of the final exam.

The basic objectives of the course as stated by the instructors—how to disassemble and assemble the guns and how to detect, diagnose, and correct malfunctioning—were assessed only informally, by casual observation of class performance.

Frederiksen and I worked for about six months developing what were termed "identification" and "performance" tests that measured the stated objectives of the course. For the identification tests, instead of the previous procedure of asking one student a few questions about the name and function of a part while the other 50 or so relaxed, each gun part was put on a cardboard (not a drawing of the gun part, but the actual part). On the cardboard were written five or more part names, including the correct one, and five or more functions, including the correct one. Fifty or so such cards, each containing five to seven part names and functions, about two or three feet apart, were distributed on tables around the room. The students arranged themselves around the room, one at each card. The students were allowed one minute at a station; then the signal was given and everyone moved one space to the right to answer the next item.

One minute was more than ample time, and the students spent a large part of the period observing the answers given by the student at the next station. On subsequent administrations, the instructions were, "Move two spaces to the right," and the move was proctored carefully. This change also necessitated two further modifications in the program. An odd number of items was required, and the class was directed to make two circuits of the room. These changes were sufficient to discourage efforts to see what other students had answered.

In the performance tests the students disassembled and assembled a rifle or an automatic pistol, a machine gun firing mechanism, or a breechblock. The proctors for the performance test had a definite list of items to watch, and they marked the student as having performed or not performed each of the listed items. The instructors pointed out that they had not been able to use performance tests, because there was not enough equipment to allow each of 50 students to have a rifle, machine gun, or breechblock. So we arranged a procedure in which enough equipment was set up for 8 or 10 students, with 8 or 10 proctors assigned to the room. Students were then called out of the written test in groups of 8 or 10, spent a half hour taking the performance test, and then returned to the room to resume the written test. Thus, at the end of a two and one-half hour test period, each student would have taken a half-hour performance test, individually supervised and marked by a proctor; taken a 45- to 50-minute identification test giving the name and function of approximately 40 items; and spent a little more than an hour on a written test.

When course grades were assigned on the basis of such an examination system, it was found that the tests of mechanical aptitude and mechanical knowledge had high validity, while the validity of the reading test dropped. This new validity pattern indicated that the students were coming closer to learning what the instructors wanted them to learn. Also, the class behavior of the students was markedly altered after the new identification and performance tests had been administered, with students spending more time in the lab disassembling and assembling the various weapons, asking other students to time them to see how long a given job took, checking with each other or the instructor on the name and function of a part, etc.

In the basic engineering school we had a similar experience. The initial validity survey showed that the Arithmetic Test (ARI) had the highest validity, and the Mechanical Knowledge Tests (Mechanical Knowledge, Mechanical—MKM; and Mechanical Knowledge, Electrical—MKE) and Mechanical Aptitude Test (MAT) had the lowest validity. The purpose of the basic engineering school was to teach the students to use various pieces of equipment, such as a lathe, drill press, and power saw, in order to make metal pieces that satisfied certain measurement specifications and could then be used as replacement parts in repair jobs. Clearly, for a course with such objectives, it was inappropriate for the ARI test to have the highest validity, and the MKM, MKE, and MAT tests to have the lowest validity. A week or so of observing classes and studying the grade records revealed what had been happening.

In the basic engineering course, one of the 12 weeks was devoted to shop arithmetic, and the grades for that week ranged from 10 or 15 to better

than 90 (mean = 83, SD about 8; see Table 1). The other 11 weeks were devoted to actual practice in the shop, using lathes, drill presses, saws, etc. The student's performance was rated by various instructors who watched him work for a while, jotted down a grade, and also carefully inspected the pieces the student turned out, measuring them and grading on closeness of agreement with initial specifications. As Table 1 shows, these shop grades had a mean of 83 or 84, but had a standard deviation of only about 2.5. Apparently, the instructors did not think that the students showed a very high performance level but were usually reluctant to give a very low mark. Also, we took a set of 30 samples made by the students and had them independently graded by several instructors. The grading of each piece was a fairly time-consuming process, involving use of squares, rulers, calipers, and so forth, to determine how closely the piece conformed to the initial specifications. A large part of the shop work grade for several weeks was determined by the marks assigned to these samples. We found that the grades for the different instructors on the set of 30 samples correlated from +.55 to −.11. Clearly, therefore, merely increasing the weighting assigned to the shop grades was not appropriate.

Nicholas Fattu worked for a year developing gauges to measure the products quickly and accurately. These gauges allowed grading of a piece within 30 seconds instead of three or four minutes. Using these gauges, the ratings of two instructors were found to correlate from .92 to .96 with each other. With the use of these gauges and identification and performance tests, the validity of the ARI dropped from highest to lowest (about .32 to .24), while the MKM, MKE, and MAT tests changed from lowest to highest (from .19 and .24 to .5 and .6; see Figure 7).

Charles Harsh worked in the Torpedoman School with similar results. Figure 8 shows that before the introduction of the identification and performance testing and grading procedures, the mechanical tests had the lowest validity (.3 to .45), while ARI had the highest validity, .64. Afterwards, the General Classification Test (GCT), Reading Test (READ), and ARI had the lowest validity (.3 to .4), while the MAT, MKM, and MKE tests had the highest validity (.45 to .62).

It is especially interesting to find that aptitude tests can indicate whether course grades are measuring relevant or unimportant aspects of the course. I had not realized before that aptitude tests could be useful in evaluating the teaching, testing, and grading for a course.

Testing Organization Reports That May Indicate Criterion Problems

In scanning reports from various testing orga-

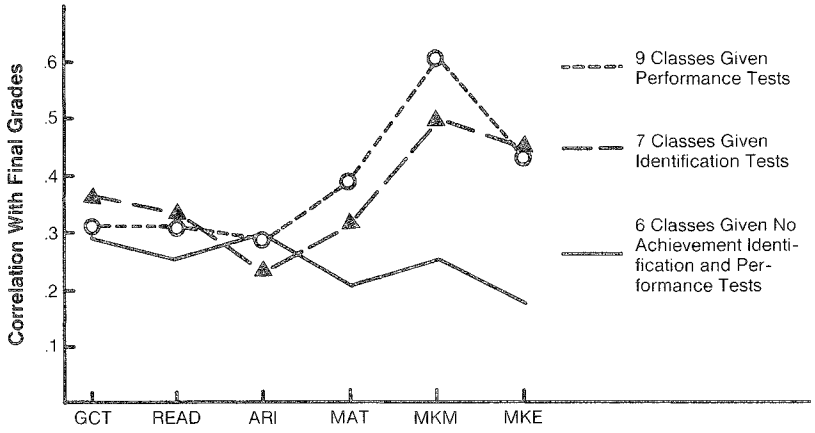
Table 1
The Relative Contribution of Each Part Grade to the Total Variance of the Composite Final Grade for Graduates of Two Classes in a Basic Engineering School

Part Grades	Class I (N = 350) M		r_{jT}	Class II (N = 340) M		r_{jT}	Relative Contribution of Each Part to Total Variance Class I	Class II
Mathematics	83.6	7.7	.86	83.4	8.5	.88	.61	.62
Mechanical Drawing	89.1	4.1	.74	87.7	4.1	.72	.28	.25
Shop	84.0	2.5	.48	83.5	2.6	.60	.11	.13

Note. From Personnel research and test development in the Bureau of Naval Personnel, by D. B. Stuit (Ed.). Copyright 1947 by the Princeton University Press. Reprinted with permission.

r_{jT} = correlation of part j with composite total T.

Figure 7
Prediction of Success in Basic Engineering School
by Use of the Basic Test Battery
(Before and After Introduction of Achievement Testing Program)
Figure 8-XV from D. B. Stuit, p. 307.



nizations, unusually low or high validities may be found. Further study might then be warranted to see if they are repeated, and if so, to investigate the particular situation further (as was done with the Navy service schools) to try to determine the reason for such validities and to take steps appropriate to correct the situation.

Some years ago, in reviewing validity coefficients for the Differential Aptitude Tests (DAT), I noticed that for one school the best predictor of grades in Latin was the clerical test (.47). For the other tests of the DAT, the correlations with Latin grades for that school ranged from a low of −.37 for mechanical reasoning through −.02 for verbal reasoning, to a high of .19 for sentences (cf. Bennett, Seashore, & Wesman, 1959, p. 48). This was not generally true for all of the schools studied. These data suggested that steps should be taken to alter the teaching and grading procedures in that school. Other studies (cf. Bennett et al., 1959, p. 79) showed that higher educational level corresponds with higher clerical ability. This may indicate an undue emphasis on clerical ability in college selection.

Some interesting results from the CEEB (1972–1974) are shown in Table 2. For females in a given school, College A, the correlation of freshman grade-point average with high school average is .44, and with the New York Regents scholarship exam, .39—low but not unusual. For males the corresponding correlations are .22 and −.02. Such an unusually low correlation is clearly not due to restriction of range, since the standard deviation for the males is greater than for females: .64 versus .53 for grade-point average, and 30.9 versus 29.9 for the Regents Exam. A closer investigation of the situation might well reveal some interesting differences in the academic or nonacademic programs for boys and girls that would help to explain these results.

Another college, College B, shows similar but perhaps less extreme differences. The correlation of freshman grade-point average with the SAT math score (SAT-M) is .34 for females and only .09 for males, despite the fact that the standard deviations for males are higher than for females—.75 to .69 for freshman grade-point average, and 8.2 versus about 8 for SAT-M. It would be interesting to check on whether or not such results are repeated in other years at these particular colleges, and to investigate possible explanations for such results.

Figure 8
Prediction of Success in Torpedoman School
by Use of the Basic Test Battery
(Before and After Introduction of Achievement Testing Program)
Figure 9-XV from D. B. Stuit, p. 308.

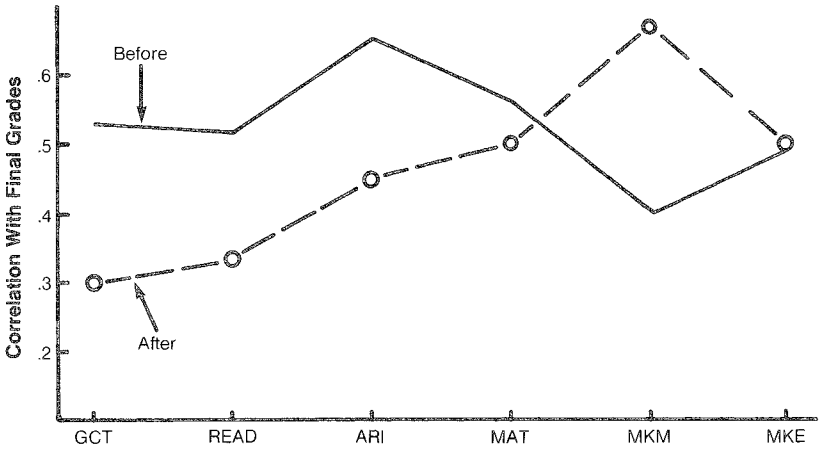


Table 2
Comparison of Grades and SAT Scores of Freshmen at Two Colleges

Variable	Mean	S.D.	r with Fr GPA	Mean	S.D.	r with Fr GPA
Freshman Students 1974, College A						
	Male (N=108)			Female (N=213)		
Criterion						
Fr GPA	2.23	.64		2.61	.53	
Predictors						
HS Ave	83.77	3.44	.225	86.72	3.63	.444
NY Reg SE	190.74	30.89	-.019	175.35	29.93	.393
Freshman Students 1972, College B						
	Male (N=85)			Female (N=200)		
Criterion						
Fr GPA	2.19	.75		2.58	.69	
Predictors						
SAT-V	44.11	7.37	.233	44.12	8.34	.329
SAT-M	48.87	8.21	.089	45.03	7.97	.342
HS Ave	2.71	.55	.419	2.94	.49	.398

Note. Adapted from tables on pages 7 and 22 of College Entrance Examination Board Reports, 1972-74.

The Problem of Testing Only for Minimum Requirements

Some validity coefficients from the Law School Admissions Test (LSAT) raise interesting questions regarding the differences in curriculum and grading procedures in law schools. Validity coefficients for predicting first-year average from LSAT (see Table 3) vary from a high of .71 to a low of .18. The standard deviations of the aptitude scores are larger year by year for school P, with validities of .18 to .21, than they are for school A with validities of .71 to .56. Furthermore, the results for 1973 and 1974 are very similar. A problem exists here regarding the curriculum and the grading procedures in these schools.

While it is true that a test only of minimum requirements would show a low validity coefficient when all the members of a class are above the minimum, such a test would then have a very small standard deviation when all students were above the minimum, and hence all would have nearly perfect scores. Such is clearly not the case for schools A and P shown in Table 3. The test variance is large for school P, and the validities are low.

As to the question of testing only for minimum requirements, an appropriate testing and grading system would include a measure of "minimum requirements"—for determining the pass-fail boundary—and would also test for achievement beyond the minimum requirements in order to give due recognition to students attaining various higher levels of performance, up to "superior achievement," the A versus B boundary, and possibly also the A+ versus A boundary. It is inappropriate for the class instructor to decide that he/she will distinguish only between "pass-fail" or "superior versus ordinary" achievement. The instructor should keep the students informed regarding their level of achievement, from failure to barely passing, to ordinary, and up to superior performance.

Various researchers at ETS have studied the reliability of grades, and their relation to other grades and to various measures of job performance (cf. Carlson & Werts, 1976; Crooks, Campbell, & Rock, 1979; Grandy, Werts, & Schabacker, 1977; Werts, Linn, & Jöreskog, 1978). Some of the results would be interesting and important to discuss with the schools and instructors involved.

Validity and Memorizing Ability

Some interesting preliminary data have been obtained at ETS using a test called Formulating Hypotheses (FH). Each item of the test is a brief description of a research study, a table, or a graph showing the results of the study, and a statement in words of the major finding. The task of the person taking the test is to write hypotheses (or "possible explanations") of that finding. The candidate is asked to write not only the hypothesis he or she thinks is most likely to be correct, but also to give other competing hypotheses that should be considered. Note that this is a free-response test—not multiple-choice. Six scores are obtained reflecting both quality and quantity of ideas. The

Table 3
Validity Coefficients (r) and Standard Deviations of LSAT Scores

	Class Entering					
	1973		1974		Combined	
School	r	S.D.	r	S.D.	r	S.D.
A	.71	78	.56	61	.62	72
N	.67	99	.51	59	.60	95
P	.18	82	.21	70	.20	76
U	.21	75	.19	57	.21	69

Note. Based on data from Barbara Pitcher (Personal Communication, 1974).

score of most interest here is one reflecting the number of hypotheses that are both unusual and of high quality.

The test was administered to students entering a medical school, and a year later the correlations of first-year grades with scores on FH and various other measures were computed (see Table 4). The best predictor of first-year grades, as might be expected, was undergraduate grade-point average (UGPA), with correlations ranging from .25 to .36. Other good predictors included scores on a biology achievement test and the verbal score on the Medical College Admissions Test. All the foregoing correlations were, of course, positive. But the best predictor, after undergraduate grade-point average, was a FH score, the one reflecting the number of hypotheses that were both unusual and of high quality. The interesting part is that these correlations were *all negative*. Two were significant at the 1% and two at the 5% level.

The first-year courses include such areas as gross anatomy, neuroanatomy, and histology. Undoubtedly, a great deal of memorization is required in such courses. It also seems that those students most willing and able to learn the names and functions of hundreds of bones, nerves, and muscles are less able (or willing) to think of and write down good hypotheses that are also unusual. Perhaps we should consider the possibility that when course grades in anatomy and histology are used as the criteria for selecting tests for admission to medical school, we may be systematically excluding some students who would be superior in creative problem-solving of the kind required, for example, in making differential diagnoses in difficult cases.

The negative validities found here for the FH scores are interesting, but they must be replicated before we take them too seriously. Attempts are now being made to obtain data at another medical school to see if similar relationships are found there. It is interesting, however, to find a situation in which negative validities are not unreasonable.

The correlation of Iowa ACT scores with college grade-point average varies for different schools from a high of .7 or .8 to below .4 (see Table 5). For the Natural Science Test, one school is above .7, while 81 of 121 schools showed correlations less than .4. It is very improbable that all these schools were teaching the same science course and grading it in the same way. Similarly, for English a few schools showed validities over .7, while for 20 of 135 schools it was below .4. It is interesting to

Table 4
Correlations of Scores on Biology Achievement Test (BAT) and Medical College Achievement Tests (MCAT), Undergraduate GPA (UGPA) in All Courses, and Scores on Formulating Hypotheses Test with First-Year Grades in Medical School (N = 80)

					Formulating Hypotheses			
		MCAT						No. of Unusual
Grades	BAT	Verb.	Sci.	UGPA	Highest Quality	No. of Resp.	Unusual Resp.	High Qual. Responses
Histology	.24	.18	.19	.36*	.08	.02	-.18	-.24*
Neuro Anat.	.24	.26	.18	.25*	.02	-.09	-.25	-.26*
Gross Anat.	.24	.27	.18	.35*	-.02	-.04	-.19	-.33*
Overall Anat.	.26	.26	.20	.36*	.04	-.05	-.24	-.31*
Class Rank Reverse	-.36*	-.36*	-.40*	-.42*	-.14	-.15	.07	.12

Note. Based on data from Norman Frederiksen (Personal Communication, 1976).
* Significant at 5% level.
* Significant at 1% level.

Table 5
ACT Research Service Frequency Distributions of Validity Correlations (r) for ACT Tests with College GPA

r Midpoint	ACT Score Engl.	Math.	Soc. Stud.	Natl. Sci.	Over-all
.80					1
.71	3	3	3	1	4
.62	23	3	11	2	22
.53	51	10	49	5	39
.44	38	21	41	32	37
Lower than .4	20	69	23	81	33
Total	135	106	127	121	136

Note. Adapted from Table 4.1 of Interpreting the 1961 ACT Research Reports (Lindquist, 1961).

note that low correlations dominate for Mathematics and Natural Science. For Mathematics, 69 of 106 schools showed correlations below .4; for Natural Science, 81 of 121 schools had correlations below .4. Such low validities for English were found for only 20 out of 135 schools, and for Social Studies for 23 out of 127. Also, for each of the tests, one or two schools showed validities above .7. It would be extremely interesting to try to ascertain the nature of the teaching, testing, and grading procedures in one or two of the high-validity schools, and also in a few of the low-validity schools, to learn what aspects of the teaching and grading procedures were responsible for such variations. We could then discuss with school authorities the extent to which the facts brought to light indicated the desirability of alterations in the teaching program and grading procedures, as well as possible alterations or additions to the testing program.

Validity and Mechanical Ability

A Junior College program tried out nine special ability tests, including Intersections, Spatial, Tool Knowledge, Mechanical Movements, and Mechanical Ability (cf. Comparative Guidance and Placement Program, 1974). Table 6 shows the results for two of these tests (Tool Knowledge and Mechanical Ability) in 80 different schools of five different types (liberal arts, science and pre-engineering, fine arts, technical science and engineering, and business). The Tool Knowledge Test showed a zero or negative validity in all schools except one. For the 14 technical science and engineering schools, there was a significant positive validity coefficient of .29 in one school, and a significant negative validity of $-.25$ in another. Sixteen schools in other categories showed negative coefficients, and the remaining 62 showed essentially zero validity coefficients.

The Mechanical Ability Test gave variable results for the fine arts schools; one school had a significant positive validity of .35, and another school a significant negative validity of $-.62$. In four of the science and engineering schools, mechanical ability showed a positive correlation with grades. In the remaining 10, the correlation was zero. Four of the business schools showed a negative validity for the Mechanical Ability Test. The remaining 18 showed essentially zero validity. For the remaining schools, three gave negative validities, and the remaining 36 essentially zero. Again, if repetition of such results in certain schools can be demonstrated, it would be very interesting to observe more closely the teaching, examining, and grading procedures to determine the reason for these differences and to consider with the school authorities whether or not changes might be desirable.

The Comparative Guidance and Placement Program (CGP) also shows a similarly wide range of

Table 6
Correlations of Scores on Tool Knowledge and Mechanical Ability Tests with Freshman Grade Average

Test and Statistic	College Parallel: Lib. Arts* (N=25)	College Parallel: Sci. & Pre-Engr. (N=14)	College Parallel: Fine Arts (N=5)	Occup.-Tech.: Sci. & Engr. (N=14)	Occup.-Tech.: Business (N=22)
Tool Knowledge					
Median	-.11	-.16	-.15	.16	-.08
Low	-.38	-.36	-.63	-.25	-.37
High	+.34	+.15	.33	+.29	+.18
No Significant					
Positive	0	0	0	1	0
Negative	5	3	3	1	5
Mechanical Ability					
Median	-.08	-.08	-.14	.21	-.03
Low	-.36	-.26	-.62	-.17	-.32
High	+.39	+.25	+.35	+.35	+.19
No. Significant					
Positive	0	0	1	4	0
Negative	2	1	1	0	4

N = Number of Correlations.
Note. Adapted from page 68 of Summary of Validity Study Results for the First Semester of the 1973-74 Academic Year (Comparative Guidance and Placement Program).

results for validity coefficients predicting the first semester of college performance, as seen in Table 7. Coefficients for the high school record range from .18 to .62, and for the CGP test from .24 to .67. Attempting to identify the school practices that produce such differences would be interesting, and might well assist the school or college to improve its teaching and grading procedures.

Course Grades as a Criterion

Goldman and Slaughter (1976) have found that the multiple correlation of SAT and high school grade-point average with grades is relatively high for individual classes such as psychology, sociology, biology, chemistry, or physics. For the total freshman class, however, the multiple correlation of SAT and high school grade-point average is low, because the general grade-point average is a composite of nonequivalent highly fluctuating parts—whereas a grade in a psychology or biology class is reasonably homogeneous. The prediction of grades in individual courses would yield more information than the prediction of grades overall. The GPAs of different students contain different mixes of courses. This kind of averaging easily lowers the validity.

Conclusions

Testing organizations should evaluate the criteria that are available for determining the validity of their tests, and, where appropriate, should point out that these criteria may well be in need of revision and improvement. Probably the testing organization should guide the revision of the criteria. It would seem that this is not done nearly as often as might be desired, especially for school grades.

Teacher-made tests should also be evaluated in terms of the objectives and the desired outcomes of instruction. Carlson's (1985) handbook on writing various types of objective items is an important initial step in this direction. Handbooks should also

Table 7
Comparative Guidance and Placement Program (CGP) Validity for All Freshmen in Each of Five Colleges' Criterion, College Performance for First Semester of Freshman Year

High School Record	CGP Plus High School Record	CGP	N
.189	.403	.403	372
.317	.437	.406	2,474
.622	.708	.666	397
.571	.656	.587	254
.241	.277	.245	632

Note. Adapted from Table I of Summary of Validity Study Results for the First Semester of the 1973-74 Academic Year (Comparative Guidance and Placement Program).

be prepared showing how to construct and grade various sorts of free-answer items. Various types of free-answer items have been used and evaluated (e.g., Frederiksen & Ward, 1978; Ward, 1982; Ward, Frederiksen, & Carlson, 1980). Ward et al. (1980) showed that the free-answer form measured ideational fluency, an ability not tested by the parallel objective form. Ward (1982) found that the same abilities were measured by both the free-answer and objective forms. Also, performance and identification items have been found to be very useful by research workers at the Center for Occupational and Professional Assessment at ETS who have developed and evaluated various types of performance and laboratory tests for a variety of occupations (see Rosenfeld & Thornton, 1978; Thornton, 1979; Thornton & Rosenfeld, 1980). The bibliography in Thornton and Rosenfeld includes reports on the development and validation of examinations for police, fire fighters, and pharmacists.

Wesman (1972) emphasized the proper attitude on alleged bias in tests:

> You don't cure malnutrition by throwing out the scale that identifies the babies who are underweight. You don't win a war by killing the messenger who brings news of defeat in a skirmish. If tests reveal that the disadvantaged have been deprived of opportunities to learn fundamental concepts, the remedy is to provide those opportunities—not do away with the source of information. If it is true that minority children do not have the motivation to learn . . . we should work toward instilling that motivation and not pretend (by neglecting to test) that they actually have learned, thus dooming them to future failure to learn.
>
> To make tests the scapegoat for the ills of the disadvantaged is not only unfair to test publishers and authors, it is unfair to a society that needs to know and to grow. . . . The remedy for the ills of society is not to dispense with diagnosis: it is to treat the ills. . . . I cannot accept the proposition that the solution to the problems of society is ignorance of facts. We must know what we are, if we are to know what we can become. (pp. 401–402)

The orientation I am urging, toward using tests to evaluate the criteria, implies an alteration in various testing procedures. For instance, there is now a tendency to remove from the test battery any test with low validities. I would suggest that in many cases such tests should be kept in the battery, just as The Psychological Corporation has

kept the clerical speed and accuracy test in its DAT battery. In general, it would probably be useful to extend test batteries to include tests not now used. Various memory tests would be an example of such an extension, including immediate memory span, rote memory, and meaningful memory. It seems that such tests probably are omitted because memorizing is considered an ability that should not be important in academic work. But if it is agreed that this is so, then it would be very important to include various kinds of memory tests in the battery, in order to learn whether there were courses for which this ability was important. On finding such courses, the instructor and tester would then have to consider whether or not memory ability was appropriately involved in a given course.

I referred to courses, rather than to grade-point averages, and feel that much more attention should be paid to prediction for various courses and to differential prediction of various specialties, such as Paul Horst emphasized at the University of Washington. Information on the abilities that show high and low validities for a given course or specialty would usually be much more informative and interesting than similar statements applied to a total academic curriculum, as illustrated in the Washington Pre-College Testing Program (see Horst, 1954; Lunneborg, 1966; Noeth, 1979).

There is also a tendency to report only validities of tests that are high, or that the tester feels should be high (Chauncey & Dobbin, 1963). Data on tests with low validities are often not reported at all. The attitude I am urging is that validities should be reported on a fairly comprehensive battery of tests covering abilities that should have low validities, as well as abilities that should have high validities. The results would be reported and interpreted, indicating that validities that should be high are high, and those that should be low are low, indicating an appropriate state of affairs. On the other hand, it could be pointed out that certain high validities should be low, and certain low validities should be high. The problem could then be discussed with the instructors involved, and their teaching and testing procedures could be examined to see if some possible changes in either might be tried.

References

Adkins, D. C. (1974). *Test construction: Development and interpretation of achievement tests* (2nd ed.). Columbus OH: Merrill.

Adkins, D. C., Primoff, E. S., McAdoo, H. L., Bridges, C. F., & Forer, B. (1947). *Construction and analysis of achievement tests*. Washington DC: U.S. Government Printing Office.

Ahmann, J. S. (1962). *Testing student achievements and aptitudes*. Washington DC: Center for Applied Research on Education.

Beggs, D. L., & Lewis, E. L. (1975). *Measurement and evaluation in the schools*. Boston: Houghton Mifflin.

Bennett, G. E., Seashore, H. G., & Wesman, A. G. (1959). *Manual for Differential Aptitude Tests*. New York: The Psychological Corporation.

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

Burt, C. (1948). *Handbook of tests, for use in schools* (2nd ed.). London: Staples.

Carlson, A. B., & Werts, C. E. (1976). *Relationships among law school predictors, law school performance, and bar examination results* (Project Report No. PR-76-26). Princeton NJ: Educational Testing Service.

Carlson, S. B. (1985). *Creative classroom testing: Ten designs for assessment and instruction*. Princeton NJ: Educational Testing Service.

Chauncey, H., & Dobbin, J. E. (1963). *Testing: Its place in education today*. New York: Harper & Row.

College Entrance Examination Board Reports, 1972–1974. Princeton NJ: Educational Testing Service, College Board Programs Division.

Comparative Guidance and Placement Program (CGP). (1974). *Summary of validity study results for the first semester of the 1973–74 academic year*. Princeton NJ: Educational Testing Service, College Board Programs Division. (Mimeographed)

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Crooks, L. A., Campbell, J. T., & Rock, D. A. (1979). *Predicting career progress of graduate students in management* (Research Report No. RR-79-15). Princeton NJ: Educational Testing Service.

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs NJ: Prentice-Hall.

Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity in scientific problem-solving. *Applied Psychological Measurement*, *2*, 1–24.

Fuess, C. M. (1950). *The College Board: Its first fifty years*. New York: Columbia University Press.

Furst, E. J. (1958). *Constructing evaluation instruments*. New York: Longmans, Green.

Goldman, R. D., & Slaughter, R. E. (1976). Why college grade-point average is difficult to predict. *Journal of Educational Psychology*, *68*, 9–14.

Grandy, J., Werts, C., & Schabacker, W. (1977, September). *Equating of ITBS and Georgia CRT Reading and Mathematics Tests*. Princeton NJ: Educational Testing Service.

Gulliksen, H. (1985). *Creating better classroom tests* (Research Memorandum No. RM 85-5). Princeton NJ: Educational Testing Service.

Hawkes, H. F., Lindquist, E. F., & Mann, C. R. (1936). *The construction and use of achievement examinations*. Boston: Houghton Mifflin.

Horst, P. (1954). A technique for the development of a differential prediction battery. *Psychological Monographs*, *68* (9, Whole No. 380).

Lindquist, E. F. (1961). *Interpreting the 1961 ACT research reports*. Iowa City IA: Science Research Associates.

Lunneborg, C. E. (1966). A research review of the Washington Pre-College Testing Program. *Journal of Educational Measurement*, *3*, 157–166.

Monroe, W. S., DeVoss, J. C., & Kelly, F. J. (1924). *Educational tests and measurements*. Boston: Houghton Mifflin.

Noeth, R. J. (1979). Washington Pre-College Test; Counselor's guide and technical summary.

Richardson, M. W., Russell, J. T., Stalnaker, J. M., & Thurstone, L. L. (1933). *Manual of examination methods*. University of Chicago, Board of Examinations.

Rinsland, H. D. (1937). *Constructing tests and grading in elementary and high school subjects*. New York: Prentice-Hall.

Rosenfeld, M., & Thornton, R. F. (1978). *The development and validation of a fire-fighter physical selection test*. Princeton NJ: Educational Testing Service.

Ruch, G. M. (1929). *The objective or new-type examination*. Glenville IL: Scott Foresman.

Stuit, D. B. (Ed.) (1947). *Personnel research and test development in the Bureau of Naval Personnel*. Princeton NJ: Princeton University Press.

Thornton, R. F. (1979, June). *The phenomena of criterion-related validity studies: The Bermuda Triangle*. Paper presented at the annual conference of the International Personnel Management Association Assessment Council, San Diego, California.

Thornton, R. F., & Rosenfeld, M. (1980). *The design and evaluation of job analysis procedures conducted for the purpose of developing content-valid occupational assessment measures*. Princeton NJ: Educational Testing Service, Center for Occupational and Professional Assessment (COPA).

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement*, *6*, 1–11.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable versions of a test. *Journal of Educational Measurement*, *17*, 11–29.

Werts, C., Linn, R. L., & Jöreskog, K. G. (1978). Reliability of college grades from longitudinal data. *Educational and Psychological Measurement*, *38*, 89–95.

Wesman, A. G. (1950). *The three-legged coefficient* (Test Service Bulletin No. 40). New York: The Psychological Corporation. (Also in *Selected writings of Alexander G. Wesman*. New York: The Psychological Corporation, 1975).

Wesman, A. G. (1972). Testing and counseling: Fact and fancy. *Measurement and Evaluation in Guidance*, *5*, 397–402. (Also in *Selected writings of Alexander G. Wesman*. New York: The Psychological Corporation, 1975).

Author's Address

Send requests for further information to Harold Gulliksen, Educational Testing Service (05R), Princeton NJ 08541, U.S.A.

Reprints

Reprints of this article may be purchased *prepaid* for $2.50 for delivery in the U.S. or $3.00 (in U.S. funds drawn directly on a U.S. bank) elsewhere, from Applied Psychological Measurement, N658 Elliott Hall, University of Minnesota, Minneapolis MN 55455-0344, U.S.A.