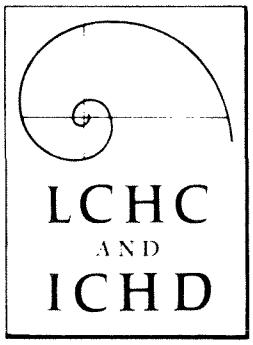
LCMall

Testing for Competence: Changing the Criterion

William S. Hall
The Rockefeller University
and
Michael Pratt
Mount Saint Vincent University



THE ROCKEFELLER UNIVERSITY

LABORATORY OF COMPARATIVE HUMAN COGNITION

AND

THE INSTITUTE FOR COMPARATIVE HUMAN DEVELOPMENT

Working Paper
No. 4
Part I

Testing for Competence: Changing the Criterion*

Introduction

Dissatisfaction with the use of intellectual aptitude tests as screening devices for nonacademic settings has led to a search for alternative measures. Disbelief in the ability of academic measures to predict performance in the nonacademic arena has led to the enactment of laws that make such a practice illegal. (See, for example Griggs vs. Duke Power Co.). What is lawful and what is the prevailing ethos in the culture are not always the same, however. Cronbach (1970), for example, has noted that most psychologists, as well as the general public, think intelligence tests tap abilities that can be responsible for job success. But the viability of this point of view has been called into question, particularly by McClelland (1973). He cites the following evidence in support of his case:

Thorndike and Hagen (1959) obtained 12,000 correlations between test scores and various measures of later occupational success on more than 10,000 respondents. They found that the number of significant correlations did not exceed what would be expected by chance.

Holland and Richards (1965) and Elton and Shevel (1969) illustrated that no consistent relationships exist between the scholastic aptitude scores of college students and their

---

*Alternatives to Standardized Testing, 1976. Symposium held at University of Pittsburgh, Pa. Unpublished.

actual accomplishment in social leadership, the arts, science, music, writing, and speech and drama.

Ghiselli's (1966, p. 121) conclusions, based on a review of 50 years of research, seem to point in the opposite direction. He found that general intelligence tests correlate .42 with trainability and .23 with proficiency across all types of jobs (correlations based on more than 10,000 cases). There are two problems with Ghiselli's data: (1) it is difficult to evaluate his conclusions, because he does not cite his sources; (2) he does not define exactly how job proficiency was measured for each of his correlations. (Measurements may have been supervisors' ratings or such indirect indicators as turnover, promotion, salary increases.) The basic problem with job-proficiency measures as these as criteria for validating ability tests is that they depend heavily on the credentials--the habits, values, background, interests, speech style--the individual brings to the job. Thus, the correlation between intelligence-test scores and job success often may be an artifact, the product of their joint association with class status or background. In general, correlations are a limited approach to issues of validity, because they provide no evidence regarding causal relations between whatever construct the test is presumed to measure and the criterion. (See McClelland, 1973, for an excellent discussion of the difficulty of evaluating "proxy" measures for job proficiency.)

Traditionally, the criterion for validating general intelligence tests has been scholastic achievement. General

intelligence tests can, therefore, be defined as measures of academic aptitude. Reliance on academic aptitude as the predicter of performance on all tasks is a direct result of the psychometric assumption of a general ability, or "g," factor. In recent years, belief in a g factor has come increasingly under attack by psychometric theoreticians.

A New Focus: The Criterion-Referenced Measures

The movement away from the assumption of g as the most important factor in human "mental" functioning, and toward a consideration of multiple traits and differential aptitudes, paved the way for the current focus on criterion-referenced testing. If we assume only a single underlying ability factor, we would not think it important to assess specific abilities for special tasks, e.g., clerical or mechanical skills. On the other hand, if we assume that humans have differential skills and that these relate differentially to various tasks, we can proceed rationally to pair those skills with different tasks, which will permit us to construct criterion-referenced measures.

Several other advantages of criterion-referenced measures can be noted. Such measures will (1) produce greater predictive validity; and (2) reduce the amount of training required in the job situation. In essence, criterion-referenced tests provide information about existing specific skills which, therefore, do not need to be taught. Tests of intellectual ability, on the other hand, select those who

have general academic skills which may not be specifically related to the job.

McClelland (1973) has stated well the case for criterion-referenced tests. He notes: "...the best testing is criterion sampling. If you want to know how well a person can drive a car (the criterion), sample his ability to do so by giving him a driver's test. If one wants to know who will make a good teacher, one should get videotapes of classrooms and find out how the behavior of good and poor teachers differ..." (p. 8). He notes further that testing calls for a revision of the role of tester, a revision that moves the tester toward behavioral analysis of both criterion measures and test instruments.

Problems in constructing a criterion-referenced measure

In constructing criterion-referenced measures, one encounters several problems. Let us consider these problems in the context of an example, specifically, designing a test to help select a patrolman from among a group of applicants. The first step is to specify the criterion, which requires agreement on the behaviors required of a good patrolman. Once this is done, a means of evaluating applicants' abilities ahead of time (the test) must be constructed. Here the form of the test becomes an important issue, because taking paper-and-pencil tests is not one of the behaviors that make up the criterion. Thus, the first problem is to define the criterion. Its corollary is formulating a method of assessment

consistent with this behavioral criterion.

Because criterion-referenced tests are designed to provide information that is directly interpretable in terms of specified performance standards, those standards must be established prior to test construction. In addition, the purpose of testing is to assess an individual's status with respect to the standards (in contrast to a comparison) with the scores of others, as in the typical norm-referenced test). Thus, securing a careful specification of the relationship between the selection measure and job performance is a second problem. But this problem can be overcome if we heed McClelland's (1973) advice to get into the "field," where we can actually analyze the various components of criterion performance. The example that follows incorporates McClelland's advice and illustrates both the procedure and several of the problems inherent in constructing criterion-referenced measures.

An example of a criterion-referenced test for job applicants

A criterion-referenced test for applicants for the position of patrolman in a northeastern state was administered three times to a total of 3,193 candidates. There were two reasons for developing such a test. One was to find an instrument that would reflect the actual tasks a patrolman must perform and would also have greater predictive validity. The second was to create a method that would give equal opportunity to both Blacks and Whites, because existing

non-job-related measures did not. The assumption underlying this two-pronged purpose was that the closer the content of a measure came to reflecting actual behaviors required on the job, the smaller the racial differences on the measure would be.

The job analysis

Development of the test began with an analysis of the specific duties required. The job analysis was structured on the basis of information gathered during individual interviews with experienced policemen. The interviewees were chosen randomly from the police ranks on the basis of number of years of service, and of ethnic and cultural background.

The interviewees were asked to list the duties or tasks of a policeman in descending order of importance. A ten-point scale was used; 1 represented the highest order of importance for that particular duty. Next, the interviewee was asked to justify his or her rank-ordering.

Altogether, 16 features were named and justified by the interviewees as follows:

1) Service to the public
2) Report-writing
3) Public relations
4) Crime prevention
5) Court appearances
6) Law enforcement

7) Apprehension of violators

8) Investigations

9) Protection of life and property

10) Patrol

11) Domestic disputes

12) Traffic control

13) First aid

14) Juvenile offenses

15) Summonses

16) Escort duties

These features were reviewed and recategorized into four tasks: 1) law enforcement; 2) public relations; 3) report writing; and 4) court appearances. An explication of the four categories was sought from a sample of senior patrolman, who were asked to specify important qualities needed to perform each task. The results can be seen in Table I.

Law enforcement. Under this category fall numbers 4, 7, 8, 10, 12, 14, 15, and 16 in the above list.

Public relations. Service to the public was found to mean a primarily public-relations function, and was, therefore, incorporated under number 1. Arbitrating domestic disputes and administering first aid (hence serving the public) were both considered to be functions of public relations and so were incorporated under that heading.

Report-writing. This duty is clearly exclusive by nature, and therefore forms a separate feature.

Court appearances. Like report-writing, court appearances fall in a special, quite separate, category.

Table 1

Tasks and Characteristics

Law enforcement	Public Relations	Report-Writing	Court Appearances
Judgment	Courtesy	Legible hand-writing	Oral fluency
Diligence	Sociability	Correct use of grammar	Correct use of grammar
Aggressiveness	Tact	Spelling ability	Good memory
Authoritativeness	Compassion	Adequate vocabulary	Adequate vocabulary
Suspiciousness	Imagination	Accurateness of expression	Confidence
Curiosity	Verbal Fluency	Conciseness of expression	
Adaptability		Concern for detail	
Impartiality			
Objectivity			

Constructing the tests

Job-related and less job-related subtests were developed for the four tasks. These are described in detail on the pages that follow.

Job-Related Subtest

Discretionary Situations:

Items were written with the intention of enabling the candidate actually to visualize himself in a potentially crucial situation, one in which any policeman might find himself during the course of the day. The activity displayed in these types of items were considered to be the crux of a patrolman's job.

Each item contained a central problem that was job-related, plausible, and required a decisive act for its solution. Ideally, the solution to the problem did not require any previous police knowledge on the part of the candidate. Four alternatives, any one of which might be feasible, were given, but only one was correct, or best, for each particular instance.

The item was intended to measure a candidate's potential for sound judgment and decision-making; consequently, a deliberate effort was made to keep the item at a low reading level for greater validity.

An example and explanation of one item is:

> You are alone in your patrol car and you have been directed to a railroad yard to investigate trespassers. As you enter the yard you observe two young boys on top of a tank car. When they see the police car they scurry down and run between the other freight cars. You drive up to the tank car and notice that the top manhold covers are open. Before going after the boys, you decide to climb to the top of the car and investigate further. Peering down into the tank car you are able to make out the figure of another boy sprawled out, seemingly unconscious.
>
> Choose the best immediate action in this instance:
>
> a. Radio for help and go after the other two boys.
>
> b. Enter the tank car and rescue the boy.
>
> *c. Radio for help and wait by the tank car.
>
> d. Drive to the control tower for help.

One can readily see that, to arrive at the correct answer, no previous police knowledge is required, even though the situation

is job-related. Ideally, the candidate should be able to place himself in such a situation and respond to the alternative which he would be most likely to choose.

Answer a is an indication of poor judgment. The policeman is investigating trespassing, not looting or a more serious offense; indeed, there is no mention of anything other than a minor infraction of the law. Also, ages of the boys must be considered; they are "young" and thus should not be treated as adults (here we get an indication of a candidate's adaptability, objectivity, and impartiality). To chase the boys through a railroad yard where there may be siding and coupling of cars or live rails would indicate a lack of insight and imagination.

Answer d would indicate a lack of diligence, aggressiveness, and authoritativeness; a candidate choosing this alternative would not have the situation under control.

Answer b, entering the tank car to rescue the boy, shows aggressiveness and authoritativeness, but not good judgment. A tank car (especially one in which there is an unconscious person) should be suspected of harboring noxious fumes. To enter unattended could be disastrous for both parties. This admittedly presupposes knowledge of enclosed vessels and fumes, but the purpose of any test is to identify the best candidate. Each candidate brings with him to the examination center a sum total of all his knowledge and experiences, which is precisely what we are trying to measure.

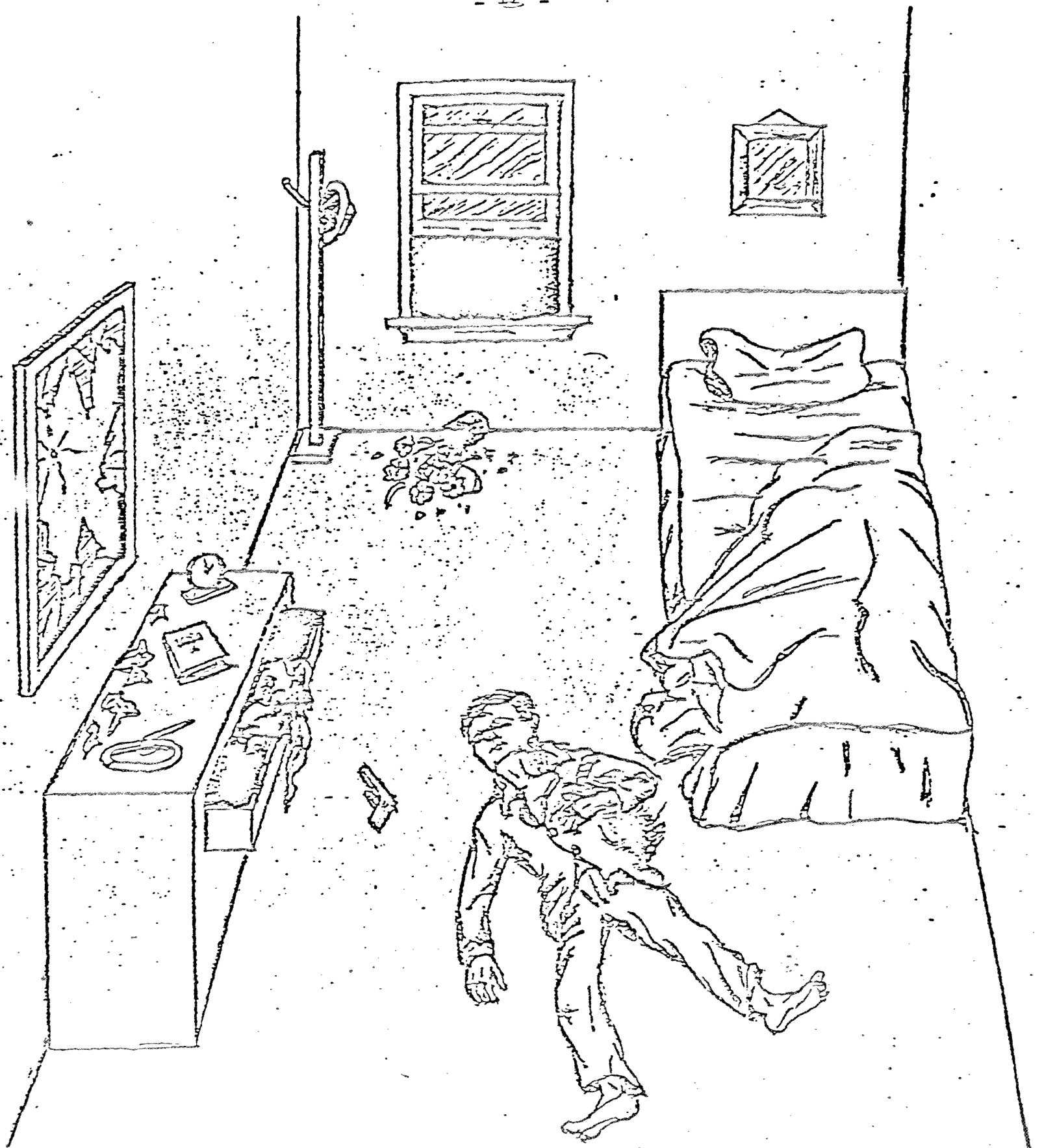
The discretionary types of items cover a magnitude of general characteristics, which often overlap into other categories. Primarily, they were designed with the intention of measuring sound judgment. However, in the kinds of situations put forth, such qualities as courtesy, sociability, tact, compassion, and imagination also can be assessed, inasmuch as choosing the correct response depends on the candidates' exhibiting one or more of these traits.

Public Relations

Originally, this subtest was designed to gauge the characteristics given in the preceding paragraph (courtesy, etc.), but because most of the variables it was designed to measure were tapped by the Discretionary Situation items, this subtest was dropped in Form 3.

Memory: This subtest was administered in the Form 3 experimental examination only. It was designed to assess attention to and careful recall of critical details, an important component skill in several of the job categories of Table I. The instructions and an example of an item used in this subtest follow:

> You will be shown a crime scene for 7 seconds. Try to remember everything you see during this period.
>
> After the crime scene is removed, check off below in Part A everything you thought you saw, whether it was in whole or in part.



Figure

Projection Scene

PART A

pajamas	table lamp	socks	clock
shoes	floor lamp	coat	rug
man	mirror	bookstand	desk
chair	picture	ashtray	potted plant
woman	slippers	book	T.V.
bed	radio	pipe	gun
dresser	dagger	cigarette	dog
night stand	cat	hat	window
child	bullet hole	cigar	wallet
couch	pen	wristwatch	coat tree

Less Job-Related Subtests

Reading Comprehension:

Reading comprehension was constructed by selecting passages of approximately 200 words from newspapers, periodicals, texts, or similar publications. When the candidate had read the passage, he/she was asked questions about it, and required to respond with the best among four alternatives. Four distinct approaches were taken in the questions:

(1) What was the main idea?

(2) What conclusion can be drawn?

(3) What statement summarizes the meaning?

(4) Which statement is correct?

The reading level of each passage was determined by the E. Fry Readability Formula. They ranged from the sixth-grade to the college level. Inasmuch as one of the academic requirements for the job of patrolman was a high-school diploma or its equivalent, the majority of the passages was kept at the tenth- to eleventh-grade level. For example:

> It would be hard to find two parts of Asia less suited to be partners. West Pakistan is part of the upland of Asia, girt by mountains, marked by vast desert areas, a plains area whose most notable characteristic is the romantic Khyber Pass. East

Pakistan is the flattest nation on earth, where an elevation of 4 feet becomes a hill. It is a mud flat where most residents have never seen a stone or rock. It is formed by two of the world's most magnificant rivers, the Ganges and the Brahmaputra, as they sprawl out at the end of their journey from the Himalayas. For hundreds of thousands of years these mighty streams have been depositing mountain silt in the Bay of Bengal, so that the underwater shelf off East Pakistan is also flat, and only a score of feet below the surface. This explains why tidal waves sometimes sweep in from the bay, inundating the flatlands and killing hundreds of thousands.

Which of the following conclusions could best be drawn from this passage?

(a) The Ganges and the Brahmaputra are also to be found in West Pakistan

*(b) The Ganges and the Brahmaputra empty into the Bay of Bengal

(c) The flattness of the country is due to the Ganges and the Brahmaputra

(d) The Himalayas are low and flat because the Ganges and the Brahmaputra originate there

Under the Fry Formula, the readability of this particular passage was gauged to be at the eighth-grade level. The content of each of the various reading passages dictated the approach taken for the queries. In this instance, it was deemed best to use the "conclusion" type of question.

Vocabulary: Fifteen nonsynonymous items displayed in a multiple-choice format comprised this subtest. No effort was made to include law-enforcement terms. Seventy-five percent of the word choices originated from Black periodical sources. They were used in a wide array of subject matter and

were chosen for their timeliness and universality.

Figure Analogies: This subtest consisted of 15 figures taken from an existing data bank. The items were similar to those found on the Wechsler coding subtest. The relationship of performance on figure analogies and patrolmanship is not known.

Number Series: In this subtest, the subject was required to detect the pattern in a series of numbers and eliminate any numbers that violated the pattern. This was dropped from forms 1 and 2 because it proved to be too easy for the applicants.

Three administrations of the aforementioned subtests were given in fairly consistent combinations (see Table II). Forms 1, 2, and 3 were administered to 1542, 438, and 1213 candidates, respectively, at different times. Approximately 13% of all candidates were Black, the rest White. These forms varied somewhat in the number of items included in the subtests; split-half reliabilities were .75-.85 for the different forms.

Reduction of racial bias in the test

A primary goal of this entire program was to develop items on which Blacks had a reasonable chance of passing without, at the same time, sacrificing the integrity of the scores achieved by the Whites. Table II presents the

differences in difficulty among the racial groups for all the subtests. The average difficulty is the proportion of items failed across all subjects in a group.

Those subtests with the most directly job-related content (discretionary situations, public relations, and memory) tended to be lowest in their discrimination against candidates on the basis of race. These results are considered to be a major contribution of the measurement enterprise being discussed here.

Table II

Summation of Average Difficulty Differences between Blacks and Whites
Forms 1, 2, and 3

Subtests	White and Black	Rank	White and Black	Rank	White and Black	Rank	Avg. Diff.
Vocabulary	.125	5	.085	2	.162	5	.124
Reading Comprehension	.089	2	.113	4	.136	4	.113
Number Series	.145	6	---	-	---	-	.145
Figure Analogies	.093	3	.124	5	.102	3	.106
Discretionary Situations	.004	1	.067	1	.032	1	.034
Public Relations	.094	4	.097	3	---	-	.096
Memory	---	-	---	-	.095	2	.095

Table IIshows that of all the subtests in the three forms, the Discretionary Situation subtest indicated the least difference between scores by Blacks and Whites. It thus ranked number 1 in terms of fairness to applicants from the two racial groups.

Public Relations and Memory tied for second place. Discretionary Situations and Memory were designed to be the most job-related of all items used in the testing program. In short, the ordering of the subtests on the degree to which they reflect racial differences in scores corresponds closely to the appropriateness of the subtests in terms of their degree of job-relatedness.

Summary

Large, unexplained differences between test scores of ethnic groups appear on standard tests of verbal skills, but do not appear, or are greatly reduced, in job-related subtests. If further evidence can substantiate these findings -- that cultural bias in item content is necessarily discrimatory with respect to certain groups--then any test of verbal skills must be job-related and job-relevent, if it is to be used for personnel selection.

We can argue defensibly that one of our objectives in devising our testing program, namely, reduction of "cultural" (racial) bias, has been accomplished. Indeed, the two most job-related of our subtests ("Discretionary Situations" and "Memory") gave Black applicants the same opportunity as White applicants to become policemen.

The issue of cultural bias in testing has a long history. Many notable attempts to address it have been reported (see, for example, Davis et al.; Cleary, 1968; Cleary and Hilton, 1968; Baehr, 1968, 1976). Our findings, though admittedly modest, contribute to this effort.

Having shown that the more job-related subtests are the least culturally biased, can we also show that they are indeed the best predicters of actual job performance? We are faced here with the critical problem of finding a way to analyze performance in the field. Several procedures for accomplishing that task are at hand. The current work employed two procedures: supervisors' ratings of patrolmen's performances and naturalistic observations. At this time, we have only correlational data on supervisors' ratings and test scores for a few subjects. We have yet to collect naturalistic observations.

Validity Study: The First Phase

Phase 1 of assessing the validity of the instrument that we have been discussing involved the use of a sample of currently employed police officers. This concurrent validity design was used both because it was practical and because the preselection sample of Blacks was inadequate. The subtests were like those described earlier in this chapter (1) Memory; (2) Discretionary Situations; (3) Reading Comprehension; (4) Vocabulary; and (5) Public Relations.

A forced choice-adjective checklist (FC-AC) was used as the criterion instrument in phase 1 of the validity study. There were four possible response adjectives for each item, two descriptive of good performance, two descriptive of poor performance.

Each of these criterion-rating forms was sent to the personnel officers in their respective police departments for distribution to superiors. It was left to the discretion of each personnel officer to select as raters those immediate supervisors who were most aware of the actual work performance of the police officer to be rated. Eighty five subjects were used in this phase of the validity study. They were drawn from 25 different police departments.

Two of the subtests (Memory and Vocabulary) were found to correlate significantly with the criterion. The memory subtest is highly job-related, whereas the vocabulary test is not. The total test was found to be associated significantly with the criterion for Blacks, but not for Whites. The reason may lie in the difference between the Black and White sample variance. The variances were significantly different at the .01 level, the distribution of Whites on the measure being much less variable than that of the Black sample.

The possibility of restriction of range must enter into any interpretation of results. In this first phase, lack of variability in the criterion scores of the Whites would have

a deflating effect on the White validity coefficient. This lack of variability may also be operating to suppress the validity coefficient for the entire sample.

This study represents only a beginning attempt at investigating the predictive functions of the present test. Finding a suitable criterion measure of performance is critical. The difficulties of doing so are illustrated in the present investigation. The use of supervisors' ratings of performance as a criterion is undesirable in several ways - especially with respect to bias and interpretation. Behavioral observations of actual performance represent a better - although more costly - alternative index. Further validity studies are being carried out, using such an observational criterion measure.

At the level of developing criterion-referenced measures, selecting a candidate for employment and prescribing an educational program have many parallels. Some can be seen in an educational example, to which we now turn.

Project TORQUE: An example of a criterion-referenced educational measure

Educational Development Corporation's Project TORQUE, a criterion-referenced approach to the assessment of children's mathematical skills, provides an educational example of programmatic test development. The TORQUE program concentrates on six skill areas in mathematics: measurement, computation, estimation, mapping and scaling, graphing, and modeling parts of the real world, using numbers.

TORQUE suggests several testing guidelines that reinforce many of the points made in the example of a patrolman's examination. The importance of reducing cultural bias (and other irrelevant factors, such as individual style) is a common theme. So is the emphasis on criterion-referencing of an individual's score and the use of actual behavioral observations to validate the test. These emphases represent major advances over traditional achievement tests, improvements closely following McClelland's (1973) recommendations for the development of new assessment strategies. The major differences between the patrolman's test and the work of Project TORQUE are in the functions of the testing. The TORQUE staff is designing tests to provide diagnostic information to teachers, so that teaching and remediation for students can be individualized and facilitated. In contrast, the police test was designed to fill a selection function, although, of course, nothing would prevent the use of this test in a program of training.

How does Project TORQUE implement its program of test development in practice? As an example of the TORQUE staff's approach to developing assessment procedure for children's skills in measurement, seventeen procedural steps are detailed. Three steps include frequent meetings with teachers who will use the tests, and discussions with children who have actually taken preliminary forms of these instruments.

The basic program calls for an initial design of "Validating Instruments," games and activities which are specifically constructed to assess skills of a particular nature.

For example, in a game called "Area-Plane," children take turns trying to cover areas with chips of three different sizes. TORQUE's developers assert that by watching a child play, an adult observer can tell in some detail what particular difficulties the child has in conceptualizing area. After extensive observations and clinical trials with children, plus ample revision, shorter "Diagnostic Achievement Tests" are written. These are also designed to assess measurement skills. The revised Validating Instruments then become the criterion against which these Diagnostic Tests must be validated. Thus, the validity of the test items "traces back to actual demonstrations of skills, rather than just expert opinion that choosing answers to items taps those skills" (Project TORQUE, 1976, p. 13). The gamelike quality of these instruments also is important in insuring student motivation.

This careful procedure, based empirically on clinical observations by the tester, and input from both those using the test (teachers) and those taking it (students) provides a much more comprehensive approach to the crucial problem of criterion measurement. The resulting assessment instruments should be of much greater diagnostic value to the teacher than standard, norm-referenced achievement tests.

References

Baehr, Melanie. "A Practitioner's View of EEOC Requirements With Special Reference to Job Analysis," Occasional Paper #37, Industrial Relations Center, The University of Chicago, September 6, 1976.

Cleary, T. Anne. "Test Bias: Prediction of Grades of Negro and White Students In Integrated Colleges," Journal of Educational Measurement, 5, #2, 115-124, 1968.

Cleary, T. Anne and Thomas L. Hilton. "An Investigation of Item Bias," Educational and Psychological Measurement, 1968, 28, 61-75.

Cronbach, L.J. Essentials of psychological testing. (3rd ed.) New York: Harper, 1970.

Davis, Allison, and Kenneth Eells. Davis-Eells Test of General Intelligence or Problem-Solving Ability Manual. Tarrytown-on-Hudson, N.Y.: World Books, 1953.

Elton, C.F., and Shevel, L.R. Who is talented? An analysis of achievement. (Res. rep. No. 31) Iowa City, Ia.: American College Testing Program, 1969.

Ghiselli, E.E. The validity of occupational aptitude tests. New York: Wiley, 1966.

Holland, J.L., and Richards, J.M., Jr. Academic and non-academic accomplishment: Correlated or uncorrelated? (Res. rep. No. 2) Iowa City, Ia.: American College Testing Program, 1965.

McClelland, David, C. "Testing for Competence Rather Than for Intelligence," American Psychologist, January, 1973, 1-14.

Project TORQUE: Tests of Reasonable Quantitative Understanding of the Environment. A New Approach to the Assessment of Children's Mathematical Competence. Project supported by a grant from Carnegie Corporation of New York: Education Development Center, Inc., 1976.

Thorndike, R.L., and Hagen, E. 10,000 careers. New York: Wiley, 1959.