Testing for Competence Rather Than for "Intelligence"

DAVID C. McCLELLAND *Harvard University* [1]

The testing movement in the United States has been a success, if one judges success by the usual American criteria of size, influence, and profitability. Intelligence and aptitude tests are used nearly everywhere by schools, colleges, and employers. It is a sign of backwardness not to have test scores in the school records of children. The Educational Testing Service alone employs about 2,000 people, annually administers Scholastic Aptitude Tests to thousands of aspirants to college, and makes enough money to support a large basic research operation. Its tests have tremendous power over the lives of young people by stamping some of them "qualified" and others "less qualified" for college work. Until recent "exceptions" were made (over the protest of some), the tests have served as a very efficient device for screening out black, Spanish-speaking, and other minority applicants to colleges. Admissions officers have protested that they take other qualities besides test achievements into account in granting admission, but careful studies by Wing and Wallach (1971) and others have shown that this is true only to a very limited degree.

Why should intelligence or aptitude tests have all this power? What justifies the use of such tests in selecting applicants for college entrance or jobs? On what assumptions is the success of the movement based? They deserve careful examination before we go on rather blindly promoting the use of tests as instruments of power over the lives of many Americans.

The key issue is obviously the *validity* of so-called intelligence tests. Their use could not be justified unless they were valid, and it is my conviction that the evidence for their validity is by no means so overwhelming as most of us, rather unthinkingly, had come to think it was. In point of fact, most of us just believed the results that the testers gave us, without subjecting them to the kind of fierce skepticism that greets, for example, the latest attempt to show that ESP exists. My objectives are to review skeptically the main lines of evidence for the validity of intelligence and aptitude tests and to draw some inferences from this review as to new lines that testing might take in the future.

Let us grant at the outset that brain-damaged or retarded people do less well on intelligence tests than other people. Wechsler (1958) initially used this criterion to validate his instrument, although it has an obvious weakness: brain-damaged people do less well on almost *any* test so that it is hard to argue that something unique called "lack of intelligence" is responsible for the deficiency in test scores. The multimethod, multitrait criterion has not been applied here.

Tests Predict Grades in School

The games people are required to play on aptitude tests are similar to the games teachers require in the classroom. In fact, many of Binet's original tests were taken from exercises that teachers used in French schools. So it is scarcely surprising that aptitude test scores are correlated highly with grades in school. The whole Scholastic Aptitude Testing movement rests its case largely on this single undeniable fact. Defenders of intelligence testing, like McNemar (1964), often seem to be suggesting that this is the only kind of validity necessary. McNemar remarked that "the manual

[1] This article contains the substance of remarks made at a public lecture given at the Educational Testing Service, Princeton, New Jersey, January 4, 1971.

Requests for reprints should be sent to David C. McClelland, Department of Psychology and Social Relations, Harvard University, William James Hall, Cambridge, Massachusetts 02138.

of the Differential Aptitude Test of the Psychological Corporation contains a staggering total of 4,096, yes I counted 'em, validity coefficients." What more could you ask for, ladies and gentlemen? It was not until I looked at the manual myself (McNemar certainly did not enlighten me) that I confirmed my suspicion that almost every one of those "validity" coefficients involved predicting grades in courses—in other words, performing on similar types of tests.

So what about grades? How valid are they as predictors? Researchers have in fact had great difficulty demonstrating that grades in school are related to any other behaviors of importance—other than doing well on aptitude tests. Yet the general public—including many psychologists and most college officials—simply has been unable to believe or accept this fact. It seems so self-evident to educators that those who do well in their classes *must* go on to do better in life that they systematically have disregarded evidence to the contrary that has been accumulating for some time. In the early 1950s, a committee of the Social Science Research Council of which I was chairman looked into the matter and concluded that while grade level attained seemed related to future measures of success in life, performance within grade was related only slightly. In other words, being a high school or college graduate gave one a credential that opened up certain higher level jobs, but the poorer students in high school or college did as well in life as the top students. As a college teacher, I found this hard to believe until I made a simple check. I took the top eight students in a class in the late 1940s at Wesleyan University where I was teaching—all straight A students—and contrasted what they were doing in the early 1960s with what eight really poor students were doing—all of whom were getting barely passing averages in college (C− or below). To my great surprise, I could not distinguish the two lists of men 15–18 years later. There were lawyers, doctors, research scientists, and college and high school teachers in both groups. The only difference I noted was that those with better grades got into better law or medical schools, but even with this supposed advantage they did not have notably more successful careers as compared with the poorer students who had had to be satisfied with "second-rate" law and medical schools at the outset. Doubtless the C− students could not get into even second-rate law and medical schools under the stricter admissions testing standards of today. Is that an advantage for society?

Such outcomes have been documented carefully by many researchers (cf. Hoyt, 1965) both in Britain (Hudson, 1960) and in the United States. Berg (1970), in a book suggestively titled *Education and Jobs: The Great Training Robbery,* has summarized studies showing that neither amount of education nor grades in school are related to vocational success as a factory worker, bank teller, or air traffic controller. Even for highly intellectual jobs like scientific researcher, Taylor, Smith, and Ghiselin (1963) have shown that superior on-the-job performance is related in no way to better grades in college. The average college grade for the top third in research success was 2.73 (about B−), and for the bottom third, 2.69 (also B−). Such facts have been known for some time. They make it abundantly clear that the testing movement is in grave danger of perpetuating a mythological meritocracy in which none of the measures of merit bears significant demonstrable validity with respect to any measures outside of the charmed circle. Psychologists used to say as a kind of an "in" joke that intelligence is what the intelligence tests measure. That seems to be uncomfortably near the whole truth and nothing but the truth. But what's funny about it, when the public took us more seriously than we did ourselves and used the tests to screen people out of opportunities for education and high-status jobs? And why call excellence at these test games intelligence?

Even further, why keep the best education for those who are already doing well at the games? This in effect is what the colleges are doing when they select from their applicants those with the highest Scholastic Aptitude Test scores. Isn't this like saying that we will coach especially those who already can play tennis well? One would think that the purpose of education is precisely to improve the performance of those who are not doing very well. So when psychologists predict on the basis of the Scholastic Aptitude Test who is most likely to do well in college, they are suggesting implicitly that these are the "best bets" to admit. But in another sense, if the colleges were interested in proving that they could educate people, high-scoring students might be poor bets because they would be less likely to show improvement in performance. To be sure, the teachers want students who will do well in their courses, but should society

allow the teachers to determine who deserves to be educated, particularly when the performance of interest to teachers bears so little relation to any other type of life performance?

Do Intelligence Tests Tap Abilities That Are Responsible for Job Success?

Most psychologists think so; certainly the general public thinks so (Cronbach, 1970, p. 300), but the evidence is a whole lot less satisfactory than one would think it ought to be to justify such confidence.

Thorndike and Hagen (1959), for instance, obtained 12,000 correlations between aptitude test scores and various measures of later occupational success on over 10,000 respondents and concluded that the number of significant correlations did not exceed what would be expected by chance. In other words, the tests were invalid. Yet psychologists go on using them, trusting that the poor validities must be due to restriction in range due to the fact that occupations do not admit individuals with lower scores. But even here it is not clear whether the characteristics required for entry are, in fact, essential to success in the field. One might suppose that finger dexterity is essential to being a dentist, and require a minimum test score for entry. Yet, it was found by Thorndike and Hagen (1959) to be related negatively to income as a dentist! Holland and Richards (1965) and Elton and Shevel (1969) have shown that no consistent relationships exist between scholastic aptitude scores in college students and their *actual accomplishments* in social leadership, the arts, science, music, writing, and speech and drama.

Yet what are we to make of Ghiselli's (1966, p. 121) conclusions, based on a review of 50 years of research, that general intelligence tests correlate .42 with trainability and .23 with proficiency across all types of jobs? Each of these correlations is based on over 10,000 cases. It is small wonder that psychologists believe intelligence tests are valid predictors of job success. Unfortunately, it is impossible to evaluate Ghiselli's conclusion, as he does not cite his sources and he does not state exactly how job proficiency was measured for each of his correlations. We can draw some conclusions from his results, however, and we can make a good guess that job proficiency often was measured by supervisors' ratings or by such indirect indicators of supervisors' opinions as turnover, promotion, salary increases, and the like.

What is interesting to observe is that intelligence test correlations with proficiency in higher status jobs are regularly higher than with proficiency in lower status jobs (Ghiselli, 1966, pp. 34, 78). Consider the fact that intelligence test scores correlate −.08 with proficiency as a canvasser or solicitor and .45 with proficiency as a stock and bond salesman. This should be a strong clue as to what intelligence tests are getting at, but most observers have overlooked it or simply assumed that it takes more general ability to be a stock and bond salesman than a canvasser. But these two jobs differ also in social status, in the language, accent, clothing, manner, and connections by education and family necessary for success in the job. The basic problem with many job proficiency measures for validating ability tests is that they depend heavily on the *credentials* the man brings to the job—the habits, values, accent, interests, etc.—that mean he is acceptable to management and to clients. Since we also know that social class background is related to getting higher ability test scores (Nuttall & Fozard, 1970), as well as to having the right personal credentials for success, *the correlation between intelligence test scores and job success often may be an artifact*, the product of their joint association with class status. Employers may have a right to select bond salesmen who have gone to the right schools because they do better, but psychologists do not have a right to argue that it is their *intelligence* that makes them more proficient in their jobs.

We know that correlation does not equal causation, but we keep forgetting it. Far too many psychologists still report average-ability test scores for high- and low-prestige occupations, inferring incorrectly that this evidence shows it takes more of this type of brains to perform a high-level than a low-level job. For instance, Jensen (1972) wrote recently:

> Can the I.Q. tell us anything of practical importance? Is it related to our commonsense notions about mental ability as we ordinarily think of it in connection with educational and occupational performance? Yes, indeed, and there is no doubt about it. . . . The I.Q. obtained after 9 or 10 years of age also predicts final adult occupational status to almost as high a degree as it predicts scholastic performance. . . . The *average* I.Q. of persons within a particular occupation is closely related to that occupation's standing in

terms of average income and the amount of prestige accorded to it by the general public [p. 9].

He certainly leaves the impression that it is "mental ability as we ordinarily think of it" that is responsible for this association between average IQ scores and job prestige. But the association can be interpreted as meaning, just as reasonably, that it takes more *pull,* more opportunity, to get the vocabulary and other habits required by those in power from incumbents of high-status positions. Careful studies that try to separate the *credential* factor from the *ability* factor in job success have been very few in number.

Ghiselli (1966) simply did not deal with the problem of what the criteria of job proficiency may mean for validating the tests. For example, he reported a correlation of .27 between intelligence test scores and proficiency as a policeman or a detective (p. 83), with no attention given to the very important issues involved in how a policeman's performance is to be evaluated. Will supervisors' ratings do? If so, it discriminates against black policemen (Baehr, Furcon, & Froemel, 1968) because white supervisors regard them as inferior. And what about the public? Shouldn't their opinion as to how they are served by the police be part of the criterion? The most recent careful review (Kent & Eisenberg, 1972) of the evidence relating ability test scores to police performance concluded that there is no stable, significant relationship. Here is concrete evidence that one must view with considerable skepticism the assumed relation of intelligence test scores to success on the job.

One other illustration may serve to warn the unwary about accepting uncritically simple statements about the role of ability, as measured by intelligence tests, in life outcomes. It is stated widely that intelligence promotes general adjustment and results in lower neuroticism. For example, Anderson (1960) reported a significant correlation between intelligence test scores obtained from boys in 1950, age 14–17, and follow-up ratings of general adjustment made five years later. Can we assume that intelligence promotes better adjustment to life as has been often claimed? It sounds reasonable until we reflect that the "intelligence" test is a test of ability to do well in school (to take academic type tests), that many of Anderson's sample were still in school or getting started on careers, and that those who are not doing well in school or getting a good first job because of it are likely to be considered poorly adjusted by themselves and others. Here the test has become part of the criterion and has introduced the correlation artificially. In case this sounds like special reasoning, consider the fact, not commented on particularly by Anderson, that the same correlation between "intelligence" test scores and adjustment in girls was an insignificant .06. Are we to conclude that intelligence does *not* promote adjustment in girls? It would seem more reasonable to argue that the particular ability tested, here associated with scholastic success, is more important to success (and hence adjustment) for boys than for girls. But this is a far cry from the careless inference that intelligence tests tap a general ability to adapt successfully to life's problems because high-IQ children (read "men") have better mental health (Jensen, 1972).

To make the point even more vividly, suppose you are a ghetto resident in the Roxbury section of Boston. To qualify for being a policeman you have to take a three-hour-long general intelligence test in which you must know the meaning of words like "quell," "pyromaniac," and "lexicon." If you do not know enough of those words or cannot play analogy games with them, you do not qualify and must be satisfied with some such job as being a janitor for which an "intelligence" test is not required yet by the Massachusetts Civil Service Commission. You, not unreasonably, feel angry, upset, and unsuccessful. Because you do not know those words, you are considered to have low intelligence, and since you consequently have to take a low-status job and are unhappy, you contribute to the celebrated correlations of low intelligence with low occupational status and poor adjustment. Psychologists should be ashamed of themselves for promoting a view of general intelligence that has encouraged such a testing program, particularly when there is no solid evidence that significantly relates performance on this type of intelligence test with performance as a policeman.

The Role of Power in Controlling Life-Outcome Criteria

Psychologists have been, until recently, incredibly naive about the role of powerful interests in controlling the criteria against which psychologists have validated their tests. Terman felt that his

studies had proved conclusively that "giftedness," as he measured it with psychological tests, was a key factor in life success. By and large, psychologists have agreed with him. Kohlberg, LaCrosse, and Ricks (1970), for instance, in a recent summary statement concluded that Terman and Oden's (1947) study "indicated the gifted were more successful occupationally, maritally, and socially than the average group, and were lower in 'morally deviant' forms of psychopathology (e.g., alcoholism, homosexuality)." Jensen (1972) agreed:

> One of the most convincing demonstrations that I.Q. is related to "real life" indicators of ability was provided in a classic study by Terman and his associates at Stanford University. . . . Terman found that for the most part these high-I.Q. children in later adulthood markedly excelled the general population on every indicator of achievement that was examined: a higher level of education completed; more scholastic honors and awards; higher occupational status; higher income; production of more articles, books, patents and other signs of creativity; more entries in *Who's Who;* a lower mortality rate; better physical and mental health; and a lower divorce rate. . . . Findings such as these establish beyond a doubt that I.Q. tests measure characteristics that are obviously of considerable importance in our present technological society. To say that the kind of ability measured by intelligence tests is irrelevant or unimportant would be tantamount to repudiating civilization as we know it [p. 9].

I do not want to repudiate civilization as we know it, or even to dismiss intelligence tests as irrelevant or unimportant, but I do want to state, as emphatically as possible, that Terman's studies do *not* demonstrate unequivocally that it is the kind of ability measured by the intelligence tests that is responsible for (i.e., causes) the greater success of the high-IQ children. Terman's studies *may* show only that the rich and powerful have more opportunities, and therefore do better in life. And if that is even possibly true, it is socially irresponsible to state that psychologists have established "beyond a doubt" that the kind of ability measured by intelligence tests is essential for high-level performance in our society. For, by current methodological standards, Terman's studies (and others like them) were naive. No attempt was made to equate for *opportunity* to be successful occupationally and socially. His gifted people clearly came from superior socioeconomic backgrounds to those he compared them with (at one point all men in California, including day laborers). He had no unequivocal evidence that it was "giftedness" (as reflected in his test scores) that was responsible for the superior performance of his group. It would be *as* legitimate (though also not proven) to conclude that sons of the rich, powerful, and educated were apt to be more successful occupationally, maritally, and socially because they had more material advantages. To make the point in another way, consider the data in Table 1, which are fairly representative of findings in this area. They were obtained by Havighurst, Bowman, Liddle, Matthews, and Pierce (1962) from a typical town in Middle America. One observes the usual strong relationship between social class and IQ and between IQ and college-going—which leads on to occupational success. The traditional interpretation of such findings is that more stupid children come from the lower classes because their parents are also stupid which explains why they are lower class. A higher proportion of children with high IQ go to college because they are more intelligent and more suited to college study. This is as it should be because IQ predicts academic success. The fact that more intelligent people going to college come more often from the upper class follows naturally because the upper classes contain more intelligent people. So the traditional argument has gone for years. It seemed all very simple and obvious to Terman and his followers.

TABLE 1

Numbers of Students in Various IQ and SES Categories (Sixth Grade) and Percentage Subsequently Going to College

IQ	Socioeconomic status			
	High	% to college	Low	% to college
High	51	71	57	23
Low	33	18	96	5

Note. $\chi^2 = 11.99$, $p < .01$, estimated tetrachoric $r = .35$, SES × IQ. (Table adapted from Havighurst et al., 1962. Copyrighted by Wiley, 1962.)

However, a closer look at Table 1 suggests another interpretation that is equally plausible, though not more required by the data than the one just given. Compare the percentages going to college in the "deviant" boxes—high socioeconomic status and low IQ versus high IQ and low socioeconomic status. It appears to be no more likely for the bright children (high IQ) from the lower classes to go to college (despite their high aptitude for it) than for the "stupid" children from the upper classes. Why is this? An obvious possibility is

that the bright but poor children do not have the money to go to college, or they do not want to go, preferring to work or do other things. In the current lingo, they are "disadvantaged" in the sense that they have not had access to the other factors (values, aspirations, money) that promote college-going in upper-class children. But now we have an alternative explanation of college-going—namely, socioeconomic status which seems to be as good a predictor of this type of success as ability. How can we claim that ability as measured by these tests is the critical factor in college-going? Very few children, even with good test-taking ability, go to college if they are from poor families. One could argue that they are victims of oppression: they do not have the opportunity or the values that permit or encourage going to college. Isn't it likely that the same oppressive forces may have prevented even more of them from learning to play school games well at all?

Belonging to the power elite (high socioeconomic status) not only helps a young man go to college and get jobs through contacts his family has, it also gives him easy access as a child to the credentials that permit him to get into certain occupations. Nowadays, those credentials include the words and word-game skills used in Scholastic Aptitude Tests. In the Middle Ages they required knowledge of Latin for the learned professions of law, medicine, and theology. Only those young men who could read and write Latin could get into those occupations, and if tests had been given in Latin, I am sure they would have shown that professionals scored higher in Latin than men in general, that sons who grew up in families where Latin was used would have an advantage in those tests compared to those in poor families where Latin was unknown, and that these men were more likely to get into the professions. But would we conclude we were dealing with a general ability factor? Many a ghetto resident must or should feel that he is in a similar position with regard to the kind of English he must learn in order to do well on tests, in school, and in occupations today in America. I was recently in Jamaica where all around me poor people were speaking an English that was almost entirely incomprehensible to me. If I insisted, they would speak patiently in a way that I could understand, but I felt like a slow-witted child. I have wondered how well I would do in Jamaican society if this kind of English were standard among the rich and powerful (which, by the way, it is not), and therefore required by them for admission into their better schools and occupations (as determined by a test administered perhaps by the Jamaican Testing Service). I would feel oppressed, not less intelligent, as the test would doubtless decide I was because I was so slow of comprehension and so ignorant of ordinary vocabulary.

When Cronbach (1970) concluded that such a test "is giving realistic information on the presence of a handicap," he is, of course, correct. But psychologists should recognize that it is those in power in a society who often decide what is a handicap. We should be a lot more cautious about accepting as ultimate criteria of ability the standards imposed by whatever group happens to be in power.

Does this mean that intelligence tests are invalid? As so often when you examine a question carefully in psychology, the answer depends on what you mean. Valid for what? Certainly they are valid for predicting who will get ahead in a number of prestige jobs where credentials are important. So is white skin: it too is a valid predictor of job success in prestige jobs. But no one would argue that white skin per se is an ability factor. Lots of the celebrated correlations between so-called intelligence test scores and success can lay no greater claim to representing an ability factor.

Valid for predicting success in school? Certainly, because school success depends on taking similar types of tests. Yet, neither the tests nor school grades seem to have much power to predict real competence in many life outcomes, aside from the advantages that credentials convey on the individuals concerned.

Are there *no* studies which show that general intelligence test scores predict competence with all of these other factors controlled? I can only assert that I have had a very hard time finding a good carefully controlled study of the problem because testers simply have not worked very hard on it: they have believed so much that they were measuring true competence that they have not bothered to try to prove that they were. Studies do exist, of course, which show significant positive correlations between special test scores and job-related skills. For example, perceptual speed scores are related to clerical proficiency. So are tests of vocabulary, immediate memory, substitution, and arithmetic. Motor ability test scores are related to proficiency as a vehicle operator (Ghiselli, 1966).

And so on. Here we are on the safe and uncontroversial ground of using tests as criterion samples. But this is a far cry from inferring that there is a general ability factor that enables a person to be more competent in anything he tries. The evidence for this general ability factor turns out to be contaminated heavily by the power of those at the top of the social hierarchy to insist that the skills they have are the ones that indicate superior adaptive capacity.

Where Do We Go from Here?

Criticisms of the testing movement are not new. The Social Science Research Council Committee on Early Identification of Talent made some of these same points nearly 15 years ago (McClelland, Baldwin, Bronfenbrenner, & Strodtbeck, 1958). But the beliefs on which the movement is based are held so firmly that such theoretical or empirical objections have had little impact up to now. The testing movement continues to grow and extend into every corner of our society. It is unlikely that it can be simply stopped, although minority groups may have the political power to stop it. For the tests are clearly discriminatory against those who have not been exposed to the culture, entrance to which is guarded by the tests. What hopefully can happen is that testers will recognize what is going on and attempt to redirect their energies in a sounder direction. The report of the special committee on testing to the College Entrance Examination Board (1970) is an important sign that changes in thinking are occurring—if only they can be implemented at a practical level. The report's gist is that a wider array of talents should be assessed for college entrance and reported as a profile to the colleges. This is a step in the right direction if everyone keeps firmly in mind that the criteria for establishing the "validity" of these new measures really ought to be not *grades in school,* but "grades in life" in the broadest theoretical and practical sense.

But now I am on the spot. Having criticized what the testing movement has been doing, I feel some obligation to suggest alternatives. How would I do things differently or better? I do not mind making suggestions, but I am well aware that some of them are as open to criticism on other grounds as the procedures I have been criticizing. So I must offer them in a spirit of considerable humility, as approaches that at least some people might be interested in pursuing who are discouraged with what we have been doing. My goal is to brainstorm a bit on how things might be different, not to present hard evidence that my proposals are better than what has been done to date. How would one test for competence, if I may use that word as a symbol for an alternative approach to traditional intelligence testing?

1. *The best testing is criterion sampling.* The point is so obvious that it would scarcely be worth mentioning, if it had not been obscured so often by psychologists like McNemar and Jensen who tout a general intelligence factor. 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. Do not give him a paper-and-pencil test for following directions, a general intelligence test, etc. As noted above, there is ample evidence that tests which sample job skills will predict proficiency on the job.

Academic skill tests are successful precisely because they involve criterion sampling for the most part. As already pointed out, the Scholastic Aptitude Test taps skills that the teacher is looking for and will give high grades for. No one could object if it had been recognized widely that this was *all* that was going on when aptitude tests were used to predict who would do well in school. Trouble started only when people assumed that these skills had some more general validity, as implied in the use of words like intelligence. Yet, even a little criterion analysis would show that there are almost no occupations or life situations that require a person to do word analogies, choose the most correct of four alternative meanings of a word, etc.

Criterion sampling means that testers have got to get out of their offices where they play endless word and paper-and-pencil games and into the field where they actually analyze performance into its components. If you want to test who will be a good policeman, go find out what a policeman does. Follow him around, make a list of his activities, and sample from that list in screening applicants. Some of the job sampling will have to be based on theory as well as practice. If policemen generally discriminate against blacks, that is clearly not part of the criterion because the law says that they must not. So include a test which shows the applicant does not discriminate. Also sample the vocabulary

he must use to communicate with the people he serves since his is a position of interpersonal influence—and not the vocabulary that men who have never been on a police beat think it is proper to know. And do not rely on supervisors' judgments of who are the better policemen because that is not, strictly speaking, job analysis but analysis of what people think involves better performance. Baehr et al. (1968), for instance, found that black policemen in Chicago who were rated high by their superiors scored high on the Deference scale of the Edwards Personal Preference Test. No such relationship appeared for white policemen. In other words, if you wanted to be considered a good cop in Chicago and you were black, you had to at least talk as if you were deferent to the white power system. Any psychologist who used this finding to pick black policemen would be guilty of improper job analysis, to put it as mildly as possible.

Criterion sampling, in short, involves both theory and practice. It requires real sophistication. Early testers knew how to do it better than later testers because they had not become so caught up in the ingrown world of "intelligence" tests that simply were validated against each other. Testers of the future must relearn how to do criterion sampling. If someone wants to know who will make a good teacher, they will have to get videotapes of classrooms, as Kounin (1970) did, and find out how the behaviors of good and poor teachers differ. To pick future businessmen, research scientists, political leaders, prospects for a happy marriage, they will have to make careful behavioral analyses of these outcomes and then find ways of sampling the adaptive behavior in advance. The task will not be easy. It will require new psychological skills not ordinarily in the repertoire of the traditional tester. What is called for is nothing less than a revision of the role itself—moving it away from word games and statistics toward behavioral analysis.

2. *Tests should be designed to reflect changes in what the individual has learned.* It is difficult, if not impossible, to find a human characteristic that cannot be modified by training or experience, whether it be an eye blink or copying Kohs' block designs. To the traditional intelligence tester this fact has been something of a nuisance because he has been searching for some unmodifiable, unfakeable index of innate mental capacity. He has reacted by trying to keep secret the way his tests are scored so that people will not learn how to do them better, and by selecting tests, scores on which are stable from one administration to the next. Stability is supposed to mean that the score reflects an innate aptitude that is unmodified by experience, but it could also mean that the test is simply insensitive to important changes in what the person knows or can do. That is, the skill involved may be so specialized, so unrelated to general experience, that even though the person has learned a lot, he performs the same in this specialized area. For example, being able to play a word game like analogies is apparently little affected by a higher education, which is not so surprising since few teachers ask their students to do analogies. Therefore, being able to do analogies is often considered a sign of some innate ability factor. Rather, it might be called an achievement so specialized that increases in general wisdom do not transfer to it and cause changes in it. And why should we be interested in such specialized skills? As we have seen, they predictably do not seem to correlate with any life-outcome criteria except those that involve similar tests or that require the credentials that a high score on the test signifies.

It seems wiser to abandon the search for pure ability factors and to select tests instead that are valid in the sense that scores on them change as the person grows in experience, wisdom, and ability to perform effectively on various tasks that life presents to him. Thus, the second principle of the new approach to testing becomes a corollary of the first. If one begins by using as tests samples of life-outcome behaviors, then one way of determining whether those tests are valid is to observe that the person's ability to perform them increases as his competence in the life-outcome behavior increases. For example, if excellence in a policeman is defined partly in terms of being evenhanded toward all minority groups, then a test of fair-mindedness (or lack of ethnocentrism) might be used to select policemen and also should reflect growth in fair-mindedness as a police recruit develops on the job. One of the hidden prejudices of psychology, borrowed from the notion of fixed inherited aptitudes, is that any trait, like racial prejudice, is unmodifiable by training. Once a bigot, always a bigot. There is no solid evidence that this trait or any other human trait cannot be changed. So it is worth insisting that a new test should be designed especially to reflect growth in the characteristic it assesses.

3. *How to improve on the characteristic tested should be made public and explicit.* Such a principle contrasts sharply with present practice in which psychologists have tried hard—backed up by the APA Ethics Committee—to keep answers to many of their tests a secret lest people practice and learn how to do better on them or fake high scores. Faking a high score is impossible if you are performing the criterion behavior, as in tests for reading, spelling, or driving a car. Faking becomes possible the more indirect the connection is between the test behavior and the criterion behavior. For example, in checking out hundreds of items for predicting flight training success, it may turn out that something like playing the piano as a boy has diagnostic validity. But no one knows exactly why: perhaps it has something to do with mechanical ability, perhaps with a social class variable, or with conscientiousness in practicing. The old-fashioned tester could not care less what the reason was as long as the item worked. But he had to be very careful about security because men who wanted to become pilots easily could report they had played the piano if they knew such an answer would help them be selected. If playing the piano actually helped people become better pilots—which no psychologist bothered to check out in World War II—then it might make some sense to make this known and encourage applicants to learn to play. That would be very like the criterion-sampling approach to testing proposed here, in which the person tested is told how to improve on the characteristic for which he will be tested.

Or to take another example, doing analogies is a task that predicts grades in school fairly well. Again no one knows quite why because schoolwork ordinarily does not involve doing analogies. So psychologists have had to be security conscious for fear that if students got hold of the analogies test answers, they might practice and become good at analogies and "fake" high aptitude. What is meant by faking here is that doing well on analogies is *not* part of the criterion behavior (getting good grades), or else it could hardly be considered faking. Rather, the test must have some indirect connection with good grades, so that doing well on it through practice destroys its predictive power: hence the high score is a "fake." The person can do analogies but that does not mean any longer that he will get better grades. Put this way, the whole procedure seems like a strange charade that testers have engaged in because they did not know what was going on, behaviorally speaking, and refused to take the trouble to find out as long as the items "worked." How much simpler it is, both theoretically and pragmatically, to make explicit to the learner what the criterion behavior is that will be tested. Then psychologist, teacher, and student can collaborate openly in trying to improve the student's score on the performance test. Certain school achievement tests, of course, follow this model. In the Iowa Test of Basic Skills, for instance, both pupil and teacher know how the pupil will be tested on spelling, reading, or arithmetic, how he should prepare for the test, how the tests will be scored, etc. What is proposed here is that *all* tests should follow this model. To do otherwise is to engage in power games with applicants over the secrecy of answers and to pretend knowledge of what lies behind correlations, which does not in fact exist.

4. *Tests should assess competencies involved in clusters of life outcomes.* If we abandon general intelligence or aptitude tests, as proposed, and move toward criterion sampling based on job analysis, there is the danger that the tests will become extremely specific to the criterion involved. For example, Project ABLE (Gagné, 1965) has identified over 50 separate skills that can be assessed for the exit level of millman apprentice (job family: woodworker and related occupations). They include skills like "measures angles," "sharpens tools and planes," and "identifies sizes and types of fasteners using gauges and charts." This approach has all of the characteristics of the new look in testing so far proposed: the tests are criterion samples; improvement in skill shows up in the tests; how to pass them is public knowledge; and both teacher and pupil can collaborate to improve test performance. However, what one ends up with is hundreds, even thousands, of specific tests for dozens of different occupations. For some purposes it may be desirable to assess competencies that are more generally useful in clusters of life outcomes, including not only occupational outcomes but social ones as well, such as leadership, interpersonal skills, etc. Project ABLE has been excellent at identifying the manual skills involved in being a service station attendant, but so far it has been unable to get a simple index of whether or not the attendant is pleasant to the customers.

Some of these competencies may be rather traditional cognitive ones involving reading, writing, and calculating skills. Others should involve what traditionally have been called personality variables, although they might better be considered competencies. Let me give some illustrations.

(*a*) *Communication skills.* Many jobs and most interpersonal situations require a person to be able to communicate accurately by word, look, or gesture just what he intends or what he wants done. Writing is one simple way to test these skills. Can the person put together words in a way that makes immediate good sense to the reader? Word-game skills do not always predict this ability, as is often assumed. I will never forget an instance of a black student applicant for graduate school at Harvard who scored in something like the fifth percentile in the Miller Analogies Test, but who obviously could write and think clearly and effectively as shown by the stories he had written as a reporter in the college paper. I could not convince my colleagues to admit him despite the fact that he had shown the criterion behavior the Analogies Test is supposed to predict. Yet if he were admitted, as a psychologist, he would be writing papers in the future, not doing analogies for his colleagues. It is amazing to me how often my colleagues say things like: "I don't care how well he can write. Just look at those test scores." Testers may shudder at this, and write public disclaimers, but what practically have they done to stop the spread of this blind faith in test scores?

In Ethiopia in 1968 we were faced with the problem of trying to find out how much English had been learned by high school students who had been taught by American Peace Corps volunteers. The usual way of doing this there, as elsewhere, is to give the student a "fill in the blanks," multiple-choice objective test to see whether the student knows the meaning of words, understands correct grammatical forms, etc. We felt that this left out the most important part of the criterion behavior: the ability to use English to communicate. So we asked students to write brief stories which we then coded objectively, not for grammatical or spelling correctness, but for complexity of thought which the student was able to express correctly in the time allotted. This gave a measure of English fluency that predictably did correlate with occupational success among Ethiopian adults and also with school success, although curiously enough it was significantly negatively related to a word-game skill (English antonyms) that more nearly approximates the usual test of English competence (Bergthold, 1969).

Important communication skills are nonverbal. When the proverbial Indian said, "White man speak with forked tongue," he doubtless meant among other things that what the white man was saying in words did not jibe with what he was doing or expressing nonverbally. The abilities to know what is going on in a social setting and to set the correct emotional tone for it are crucial life-outcome criteria. Newmeyer (1970), for instance, has found a way to measure success at enacting certain emotions so that others receive them correctly and to measure success at receiving the correct emotions over various enactors. He found that black boys at a certain age were consistently better than white boys at this particular kind of communication skill, which is a far more crucial type of criterion behavior than most paper-and-pencil tests sample.

(*b*) *Patience,* or response delay as psychologists would call it, is a human characteristic that seems essential for many life outcomes. For instance, it is desirable for many service occupations where clients' needs and demands can be irritating. It would seem particularly desirable in a policeman who has the power and authority to do great damage to people who irritate him. Kagan, Pearson, and Welch (1966) have shown that it is an easily measured human characteristic that is relatively stable over time and can be taught directly.

(*c*) *Moderate goal setting* is important in achievement-related games, as I have explained fully elsewhere (McClelland, 1961). In most life situations, it is distinctly preferable to setting goals either too high or too low, which leads more often to failure. Many performance situations have been devised which measure the tendency to set moderate, achievable goals and help the person learn how to set more realistic goals in the future (Alschuler, Tabor, & McIntyre, 1970; McClelland & Winter, 1969).

(*d*) *Ego development.* Many scholars (see Erikson, 1950; Loevinger, 1970; White, 1959) have reasoned that there is a general kind of competence which develops with age and to a higher level in some people than in others. Costa (1971a) recently has developed a Thematic Apperception Test code for ego development which appears to have many of the aspects sought in the new measurement

direction proposed here. The thought characteristics sampled represent criterion behavior in the sense that at Stage 1, for example, the person is thinking at a passive conformist level, whereas at Stage 4, he represents people in his stories as taking initiative on behalf of others (a more developed competency). The score on this measure predicts very well which junior or high school students will be perceived by their teachers as more competent (even when correlations with intelligence and grade performance are removed), and furthermore a special kind of education in junior high school moves students up the ego development scale significantly. That is, training designed to develop a sense of initiative produced results that were reflected sensitively in this score. Pupils and teachers can collaborate in increasing this kind of thinking which ought to prepare students for competent action in many spheres of life.

5. *Tests should involve operant as well as respondent behavior.* One of the greatest weaknesses of nearly all existing tests is that they structure the situation in advance and demand a response of a certain kind from the test taker. They are aimed at assessing the capacity of a person to make a certain kind of response or choice. But life outside of tests seldom presents the individual with such clearly defined alternatives as "Which dog is most likely to bite?" or "Complete the following number series: 1 3 6 10 15 __," or "Check the word which is most similar in meaning to lexicon . . . ." If we refer to these latter behaviors as respondents in the sense that the stimulus situation clearly is designed to evoke a particular kind of response, then life is much more apt to be characterized by operant responses in the sense that the individual spontaneously makes a response in the absence of a very clearly defined stimulus. This fact probably explains why most existing tests do not predict life-outcome behaviors. Respondents generally do not predict operants. To use a crude example, a psychologist might assess individual differences in the *capacity* to drink beer, but if he used this measure to predict actual beer consumption over time, the chances are that the relationship would be very low. How much beer a person can drink is not related closely to how much he does drink.

Testers generally have used respondent behaviors to save time in scoring answers and to get higher test–retest reliability. That is, the person is more likely to give the same response in a highly structured situation than in an unstructured one that allows him to emit any behavior. Yet, slavishly pursuing these goals has led to important lacks in validity of the tests because life simply is not that structured, and often does not permit one to choose between defined-in-advance responses. The n Achievement measure, which is an operant in the sense that the subject emits responses (tells stories) under only very vague instructions, has predicted over a 12–14-year period in three different samples those who will drift into entrepreneurial business occupations (McClelland, 1965). Here an operant is predicting an operant—the tendency to think spontaneously about doing better all the time predicts a series of spontaneous acts over time which leads the individual into an entrepreneurial occupation. But predicting from operants to respondents or vice versa does not work, at least for men (McClelland, 1966). The n Achievement score is not related to grades or academic test scores (respondent measures), nor do grades relate to entering entrepreneurial occupations (see McClelland, 1961).

Even within fairly structured test situations it is possible to allow for more operant behavior than has been the usual practice. Not long ago we tried to find an existing performance test on which a person with high n Achievement ought to do well because such a test might be a useful substitute for the Thematic Apperception Test storytelling measure in certain situations. Theoretically, such a test should permit operant behavior in which the individual generates a lot of alternatives for solving a problem in search of the most efficient solution. But to our surprise we could find no such test. Tests of divergent thinking existed that counted the number of operants (e.g., original uses for a paper clip) an individual could come up with, but they did not require the person to find the best alternative. Most other tests simply required the person to find the one correct answer the test maker had built into the item. What was needed were test items to which there were many correct answers, among which one was better than others in terms of some criteria of efficiency that the person would have to apply. This task seemed more lifelike to us and certainly more like the type of behavior characteristic of people with high n Achievement. So we invented an Airlines Scheduling Test (Bergthold, 1969) in which the person is faced

with a number of problems of getting a passenger from City A to City B by such and such a time at minimum expenditure in time, energy, money, and discomfort. From schedules provided, several alternative routes and connections can be generated (if the test taker is energetic enough to think them up) that will solve the problem, but one is clearly the most efficient. The test has promise in that it correlates with the n Achievement score at a low level. But the main point is that it requires more lifelike operant behavior in generating alternative solutions and therefore it should have more predictive power to a variety of situations in which what the person is expected to do is not so highly structured as in standard respondent tests.

6. *Tests should sample operant thought patterns to get maximum generalizability to various action outcomes.* As noted already, the movement toward defining behavioral objectives in occupational testing can lead to great specificity and huge inventories of small skills that have little general predictive power. One way to get around this problem is to focus on defining thought codes because, almost by definition, they have a wider range of applicability to a variety of action possibilities. That is, they represent a higher order of behavioral abstraction than any given act itself which has not the capacity to stand for other acts the way a word does. And in empirical fact this is the way it has worked out. The n Achievement score—an operant thought measure—has many action correlates from goal setting and occupational styles to color and time-span preferences (McClelland, 1961) which individually have little power as "actones" to predict each other. A more recent example is provided by an operant thought measure of power motivation which has very low positive correlations with four action characteristics: drinking, gambling, accumulating prestige supplies, and confessing to having many aggressive impulses that are not acted on (McClelland, Davis, Kalin, & Wanner, 1972). These action characteristics are completely unrelated to each other so that they would be unlikely to come out on the same dimension in a factor analysis. But what is particularly interesting is that they appear to be alternative outlets for the power drive because the power motivation score correlates much higher with the maximum expression of *any one* of these alternatives than it does with any one alone or with the sum of standard scores on all of them. The thought characteristic—here the desire to "have impact," to make a big splash—is the higher order abstraction that gives the test predictive power for alternative ways of making a big splash in action—by gambling, drinking, etc. The tester of the future is likely to get farther in finding generalizable competencies of characteristics across life outcomes if he starts by focusing on thought patterns rather than by trying to infer what thoughts must lie behind the clusters of action that come out in various factors in the traditional trait analysis.

However, I have been arguing for this approach for over 20 years, and as far as I can see, the testing movement has been affected little by my eloquence. Why? There are lots of reasons: People keep insisting that the n Achievement score is invalid because it will not predict grades in school—which is ironic since it was designed precisely to predict life outcomes and not grades in school. Or they argue it does not predict all types of achievement (Klinger, 1966)—when, of course, it is not supposed to, on theoretical grounds. But the practical problem (outside the tedium of content coding) is the unreliability of operant thought measures. Many of them are unreliable, though not all. Costa's (1971b) ego development score has a test–retest stability coefficient over a year of .66, $N = 223$. Unreliability is a fatal defect if the goal of testing is to *select* people, let us say, with high n Achievement. For rejected applicants could argue that they had been excluded improperly or that they might have high scores the next time they took the test, and the psychologist would have no good defense. One could just imagine beleaguered psychologists trying to defend themselves against irate parents whose children had not gotten into a preferred college because their n Achievement scores were too low.

But the emphasis in the new testing movement should be as much on evaluating educational progress as it is on identifying fixed characteristics for selection purposes. The operant thought measures are certainly reliable enough for the former objective. The educator can use them to assess whether a certain class or an innovative approach to teaching has tended, on the average, to promote ego development in thought as assessed by Costa's measure. The educator does not care *which particular child* is high in the measure since he does not plan to use the measure to select the child for special treatment. So its unreliability does not

matter. He, as an administrator, can use the test information to decide whether the goals of the school are being forwarded by one educational approach or another. In a sense, the very unreliability of the thought measures may be a virtue if they encourage educators to stop thinking only about selection and start thinking more about evaluating educational progress.

Does this mean that test reliability is always unimportant? Not at all. Sometimes it will be important to diagnose deficiencies reliably that are to be made up. On other occasions tests will have to be used to pick out those most likely to be able to do a particular job well. So something will have to be done about reliability. Thus, a man with a high n Achievement score is a better bet for a sales job than a man with a low n Achievement score, but the measure of n Achievement from content coding of thought samples is not very defensible for selection purposes because it is unreliable. In this instance, the thought code can be used as the criterion against which more reliable performance measures can be validated. For example, the Airlines Scheduling Test score is reliable, and if it turns out to be related consistently to the n Achievement score based on thought sampling, it can be used as a substitute for the latter in selection. In fact, the thought codes can be considered devices for finding the clusters of action patterns that can be measured more reliably to get indexes of various competency domains central to various life outcomes. For example, if it turns out that an elevated socialized power (s Power) score (McClelland et al., 1972) characterizes successful policemen more than unsuccessful ones—as would be expected—then the action correlates of socialized power, such as capacity to lead or be influential in social groups, can be used to select potentially good policemen. The s Power score itself could not be so used because it is unreliable and "fakeable" if you learn the scoring system, but it is essential as a validating criterion for more reliable measures because its wide network of empirical and theoretical relationships helps find the action characteristics that will be useful for selection purposes.

While the six principles just enumerated for the new testing movement may affect occupational testing, the fact remains that testing has had its greatest impact in the schools and currently is doing the worst damage in that area by falsely leading people to believe that doing well in school means that people are more competent and therefore more likely to do well in life because of some real ability factor. Concretely, what would an organization like the Educational Testing Service do differently if it were to take these six principles into account? As a start, it might have to drop the term intelligence from its vocabulary and speak of scholastic achievement tests that are more or less content specific. The non-content-specific achievements (formerly called "aptitudes") do predict test-taking and symbol manipulation competencies, and these competencies are central to certain life-outcome criteria—like making up tests for others to pass or being proficient as a clerk (Ghiselli, 1966). But it is a serious practical and theoretical error to label them general intelligence, on the basis of evidence now available.

Once the innate intelligence philosophy is discarded, it becomes apparent that the role of such a testing service is to report to schools a profile of scholastic and nonscholastic achievements in a number of different areas. Then, in the case of selection, it is for the college to decide whether it has the educational programs that will promote growth in given areas of low performance. If performance is already high, say in mathematics, then the college probably can produce little improvement in that area and should ask itself in what other areas it can educate such a student, as shown by his lower levels of accomplishment at the outset. The profile particularly should include measures of such general characteristics as ego development or moral development (Kohlberg & Turiel, 1971) based on thought samples, because these general competencies ought to be improved by higher educational systems anyway.

The profile of achievements should be reported not only at entrance but at various points throughout the schooling to give teachers, administrators, and students feedback on whether growth in desired characteristics actually is occurring. Test results then become a device for helping students and teachers redesign the teaching–learning process to obtain mutually agreed-on objectives. Only then will educational testing turn from the sentencing procedure it now is into the genuine service it purports to be.

REFERENCES

ALSCHULER, A. S., TABOR, D., & MCINTYRE, J. *Teaching achievement motivation.* Middletown, Conn.: Educational Ventures, 1970.

ANDERSON, J. E. The prediction of adjustment over time. In I. Iscoe & H. Stevenson (Eds.), *Personality development in children*. Austin: University of Texas Press, 1960.

BAEHR, M. E., FURCON, J. E., & FROEMEL, E. C. *Psychological assessment of patrolman gratifications in relation to field performance*. Washington, D.C.: United States Government Printing Office, 1968.

BERG, I. *Education and jobs: The great training robbery*. New York: Praeger, 1970.

BERGTHOLD, G. D. The impact of Peace Corps teachers on students in Ethiopia. Unpublished doctoral dissertation, Harvard University, 1969.

COLLEGE ENTRANCE EXAMINATION BOARD. *Report, Special Committee on Testing*. Princeton: Educational Testing Service, 1970.

COSTA, P. Introduction to the Costa ego development manual. Department of Social Relations, Harvard University, 1971. (Mimeo) (a)

COSTA, P. Working papers on ego development validation research. Harvard University, 1971. (Mimeo) (b)

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

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

ERIKSON, E. H. *Childhood and society*. New York: Norton, 1950.

GAGNÉ, R. M. *The conditions of learning*. New York: Holt, Rinehart & Winston, 1965.

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

HAVIGHURST, R. J., BOWMAN, P. H., LIDDLE, G. P., MATTHEWS, C. V., & PIERCE, J. V. *Growing up in River City*. New York: Wiley, 1962.

HOLLAND, J. L., & RICHARDS, J. M., JR. *Academic and nonacademic accomplishment: Correlated or uncorrelated?* (Res. rep. No. 2) Iowa City, Ia.: American College Testing Program, 1965.

HOYT, D. P. *The relationship between college grades and adult achievement, a review of the literature*. (Res. rep. No. 7) Iowa City, Ia.: American College Testing Program, 1965.

HUDSON, L. Degree class and attainment in scientific research. *British Journal of Psychology*, 1960, **51**, 67–73.

JENSEN, A. R. The heritability of intelligence. *Saturday Evening Post*, 1972, **244**(2), 9, 12, 149.

KAGAN, J., PEARSON, L., & WELCH, L. The modifiability of an impulsive tempo. *Journal of Educational Psychology*, 1966, **57**, 359–365.

KENT, D. A., & EISENBERG, T. The selection and promotion of police officers. *The Police Chief*, 1972, February, 20–29.

KLINGER, E. Fantasy need achievement as a motivational construct. *Psychological Bulletin*, 1966, **66**, 291–308.

KOHLBERG, L., LACROSSE, J., & RICKS, D. The predictability of adult mental health from childhood behavior. In B. Wolman (Ed.), *Handbook of child psychopathology*. New York: McGraw-Hill, 1970.

KOHLBERG, L., & TURIEL, E. *Moralization research, the cognitive development approach*. New York: Holt, Rinehart & Winston, 1971.

KOUNIN, J. S. *Discipline and group management in classrooms*. New York: Holt, Rinehart & Winston, 1970.

LOEVINGER, J., & WESSLER, R. *Measuring ego development*. San Francisco: Jossey-Bass, 1970. 2 vols.

MCCLELLAND, D. C. *The achieving society*. New York: Van Nostrand-Rheinhold, 1961.

MCCLELLAND, D. C. *N* Achievement and entrepreneurship: A longitudinal study. *Journal of Personality and Social Psychology*, 1965, **1**, 398–392.

MCCLELLAND, D. C. Longitudinal trends in the relation of thought to action. *Journal of Consulting Psychology*, 1966, **30**, 479–483.

MCCLELLAND, D. C. Education for competence. In H. Heckhausen & W. Edelstein (Eds.), *Proceedings of the 1971 FOLEB Conference*. Berlin, Germany: Institut für Bildungsforschung in der Max-Planck-Gesellschaft, in press.

MCCLELLAND, D. C., BALDWIN, A. L., BRONFENBRENNER, U., & STRODTBECK, F. L. *Talent and society*. Princeton: Van Nostrand, 1958.

MCCLELLAND, D. C., DAVIS, W. N., KALIN, R., & WANNER, H. E. *The drinking man*. New York: Free Press, 1972.

MCCLELLAND, D. C., & WINTER, D. G. *Motivating economic achievement*. New York: Free Press, 1969.

MCNEMAR, Q. Lost: Our intelligence? Why? *American Psychologist*, 1964, **19**, 871–882.

NEWMEYER, J. A. Creativity and non-verbal communication in pre-adolescent white and black children. Unpublished doctoral dissertation, Harvard University, 1970.

NUTTALL, R. L., & FOZARD, T. L. Age, socioeconomic status and human abilities. *Aging and Human Development*, 1970, **1**, 161–169.

TAYLOR, C., SMITH, W. R., & GHISELIN, B. The creative and other contributions of one sample of research scientists. In C. W. Taylor & F. Barron (Eds.), *Scientific creativity: Its recognition and development*. New York: Wiley, 1963.

TERMAN, L. M., & ODEN, M. H. *The gifted child grows up*. Stanford: Stanford University Press, 1947.

THORNDIKE, R. L., & HAGEN, E. *10,000 careers*. New York: Wiley, 1959.

WECHSLER, D. *The measurement and appraisal of adult intelligence*. (4th ed.) Baltimore: Williams & Wilkins, 1958.

WHITE, R. W. Motivation reconsidered: The concept of competence. *Psychological Review*, 1959, **66**, 297–333.

WING, C. W., JR., & WALLACH, M. A. *College admissions and the psychology of talent*. New York: Holt, Rinehart & Winston, 1971.