ROBERT L. THORNDIKE, Ph. D.
Professor of Education, Columbia University, New York.

TESTS AS LONG RANGE PREDICTORS OF VOCATIONAL CAREERS

An address before the Puerto Rico Psychological Association, on April 5, 1956.

It is just about forty years ago that I took my first intelligence test. I can still remember being dragged down, sleepy and mildly protesting, to serve as the guinea pig in a demonstration of what I now identify as the Stanford-Binet before a graduate students' club at Teachers College. A great many people have been given aptitude tests in those forty years. Nobody knows how many, but 40 million would certainly be a conservative estimate for the United States alone. If we start counting with the million or two of the First World War, count in all the uncomplaining school children of the period between the wars, add on the 14 million of the Second World War, and continue with the educational, military and industrial testing of the post-war period, the total is certainly quite astronomical.

What does all this aptitude testing add up to? What do we know with assurance about the significance of aptitude test results in the broad sweep of individual careers? What can we say about the significance of aptitude measures for vocational choice and vocational success over the span of the individual's life?

I submit that as far as any answer to this broad question is concerned, all our testing to date tells us very little. Of course, most of the millions of tests were given to serve immediate practical ends, or to serve no ends at all. They were given in order to decide whether to put Johnny in the fast or slow section, or they were given because a slick test salesman persuaded the superintendent of schools that any *really* up and coming school system gave intelligence tests. (Having given them, the schools then filed away the results and went their accustomed ways in peace.) This testing has left no mark on our accumulation of scientific knowledge.

But even the studies that have attempted to explore the relationships of test results to occupational choice and success have left us far from enlightened.

We have, for example, certain analyses of the general ability tests given during World Wars I and II that show the average score of men in different occupations. The most extensive and recent is Stewart's article in *Occupations*, in which data on the A.G.C.T. are given for perhaps a hundred different occupational groups. But this is testing after the fact of a limited and possibly badly biased sample of those already in the occupation and testing with respect to a single undifferentiated measure.

We have my esteemed father's study carried out from 1922 to 1930 of 2000 New York City 8th grade pupils, tested with the sorts of tests available in the early 1920's, and followed up for eight years to the age of about 22. But this group was inadequate in size, the tests were limited in variety and quality, the group was followed only into the beginnings of their working lives, and that beginning corresponded with the severe depres-

sion of the 30's. The essentially negative results for the tests may reflect these factors.

We have the validation studies reported by the Employment Service for the General Aptitude Test Battery. But these are typically based upon small groups of individuals in one or two specific companies, who were already working on the job and were tested after employment, rather than as job applicants.

I am not aware of any research that combines the following four characteristics:

1. Each individual was tested with a really comprehensive aptitude test battery that yielded scores for most or all of the major dimensions of ability with which a guidance counselor would be concerned.

2. The testing was carried out with young people prior to their actual employment.

3. The persons tested were followed over a long enough period of time to permit them to become established in their jobs, achieve a stable vocational choice, and demonstrate their effectiveness in their work.

4. A large enough group was studied to yield adequate numbers in each of a number of occupational specialities.

I am now engaged in a study that approaches these ideals, and it is this that I want to tell you about today.

I imagine that there is no need to convince a group such as this of the importance of such a study. We are ready at the drop of a hat to give tests as part of a program of counseling or in the process of screening job applicants. We glibly interpret the meaning of the test scores to the counselees or the potential employer. It is our responsibility to seek evidence of the extent to which the predictions that we make, based on test scores, will in fact hold up in the later careers of those who show particular ability patterns.

In order to describe my project and my findings to you, I must go back to 1941. That was a fateful year that many of you still remember quite vividly. The pace of European war was stepping up, and so was the tempo of United States preparation. December 7th brought Pearl Harbor and our sudden complete involvement.

In July of that year, General David Grant, Surgeon of what was then the Army Air Corps, anticipating that the program of training for Air Force pilots, navigators and bombardiers would increase enormously, and recognizing the importance of maintaining and even improving the quality of personnel under an expanded and accelerated training program, had commissioned in the Air Force the first of a large group of psychologists, who were to set up a classification program for flying personnel. From that nucleus there developed the Air Force Aviation Psychology Program, under the direction of then Colonel, now Dr. John C. Flanagan, with a strength of several hundred officers and enlisted men in a network of research and testing stations extending across the country.

An extensive testing program was set up, consisting of a full day of printed group tests and a half day of individually administered apparatus tests. The test battery was gradually improved during the war, until at the end the procedures were such that a pilot in the top scoring group had about five times as many chances of getting through pilot training as one in the bottom scoring group. Results with navigators were at least as good, though we never were able to claim complete success in spotting the good bombardiers. (Perhaps one reason was that the man who hit the bull's-eye one day was likely to completely miss the target the next). With pilots costing $25,000 apiece to train even then, the program paid for itself many times over in reduced training costs alone, to say nothing of gains in effectiveness of the final product.

That much is history. The interesting thing from the present standpoint is that when the smoke of battle died down we found that we had given our test battery to over half a million

men. Our record system had been such that all the test records were on IBM punch cards, and these cards have been preserved to this day. There exists in San Antonio, Texas, a test file containing detailed information on the aptitudes of over half a million men. A pool of basic data such as this has never existed before in the history of the world. The material is absolutely unique.

About four years ago now, as an adjunct to some contract research we were doing for the Air Force, we embarked upon an initial pilot study to see how well this material could be put to work to serve our civilian needs for information about what sorts of people go into, and what sorts of people succeeded in what sorts of jobs. Funds available to us in our Air Force contract limited us to a sample of 1500 cases. These 1500 we undertook to follow up. We wanted to find out what each man was doing, and to get what evidence we could of how successful he was in that job, whatever it might be.

For the present I won't go into the details of how we traced the men down. Primarily, we relied upon military and Veterans Administration records. We are grateful indeed for the cooperation we got from both these sources. It is, I think, of interest to report what success we had in locating this initial group. The results were essentially as follows:

Of the 1500, approximately

75 had died, either in military service or subsequently;

140 were at the time still on active duty with the Air Force;

25 were lost, as far as finding their records was concerned;

200 were cases that we were not able to complete within the time limits of the study;

240 failed to reply to any of the three inquiries that we believe reached their correct address;

750 filled out and returned the questionnaire that we sent them;

70 were reached by interview.

Counting those who had died, those who were in the Air Force, those who were reached by interview, and those who were reached by questionnaire, we found out what more than 1,000 of the 1,500 men were doing. With more time to work on the remainder, and funds for more interviewing, we could have increased the returns still further.

These men had been given the battery of aircrew tests in 1943. At that time they were young men 18 to 26. Most of them (about 95 percent) had at least completed high school, and about two-fifths had completed college. As a group, they were above average in ability, because they had been pre-screened both by a uniform screening test and by whatever self-selection restrains a man from applying for training where he knows he is likely to be rejected. Their abilities were perhaps roughly comparable to those of freshmen and sophomores in a State University. This select character of the group must be borne in mind as we look at some of the results presently.

What were these men doing ten years later? The answer is: Literally, everything under the sun. They ranged from accountants to welders, from actors to writers. Among the eight hundred and some returns, we had a chiropractor and a funeral director, a paper hanger and a locker room attendant. We even had one man who had been picked up for signing other people's names to checks and who was serving time in the state penitentiary for forgery. (Incidentally, we were quite gratified, on looking up the test records of our forger, to find that he was conspicuously low on a test of finger dexterity. If he had come to us for vocational guidance, we could have pointed out to him that forgery was clearly not the line of work for him).

All told, our men fell in some 150 Census occupational categories, many of them represented by only one or two men. For the moment our attention must be centered on the few oc-

cupational groupings in which we had somewhat more adequate representation. These are such groups as accountants, engineers, business managers and executives, salesmen, machinists, insurance agents, foremen, or machine operatives. We may ask in each case: What were the men who went into this occupation like? In what ways did those who were most successful differ from the others?

Though the tests given back in 1943 were being used to pick men for the specific military jobs of pilot, navigator, and bombardier, they were in many cases much like the aptitude tests that we use in guiding men into or selecting men for civilian occupations. Thus, they included tests of ability to read with understanding, ability to reason out numerical problems, ability to understand mechanical devices, ability to perceive details rapidly and accurately, ability to visualize spatial relations, ability to coordinate the activities of the two hands, ability to move the fingers rapidly and accurately, and a number of others. Though scores were available for each man on 16 tests, I shall refer to only eight of these that can be made fairly meaningful to you with a brief description.

First, let us take one particular occupational group, and see what they were like on certain of these eight tests. Among the 800 for whom we had information there were 29 men whose present occupation could be classified as accountant or auditor. In the first slide we show you how these men did on an *arithmetic reasoning* test. This is the familiar kind of test, made up of verbally stated problems (e. g., A plane that flies 150 miles an hour uses 75 gallons of gas an hour. How much gas does it take to fly 500 miles?) We have divided the total group of 1500 that we started with up into thirds. Remembering that our total group was already quite select, in comparison with the total population, we have called the bottom third "average or below". The middle third is labeled "above average", the upper third is called "superior". These labels provide convenient ways of referring to thirds of our group.

Of our 29 accountants, 15 (or 52 percent) are superior on

arithmetic reasoning. Eleven (or 38 percent) fall in the "above average" category, while only 3 (or 10 percent) are average or below. By contrast look at slide 2, which shows *mechanical principles* scores. The *mechanical principles test* presents pictures of mechanical situations and requires the examinee to select the solution that embodies the correct mechanical principle. For example, he must indicate which of several systems of bracing will give the strongest roof for an airplane hangar. In the "superior" group we find 9, in the "above average" group 6, and in the "average or below" group 14. If anything, accountants are more frequent in the lower groups.

What about *successful* accountants. We have defined "success" to a rough first approximation as the income the individual reports receiving for his work. We recognize that income is only one measure of success. We recognize, further, that individuals will be somewhat inaccurate in reporting that income. However, we haven't yet figured out any good reason for expecting the people who score high on our tests to lie about their incomes more than the low scorers. So we have taken reported income as one admittedly rough, but presumably unbiased, measure of success.

Slide 3 shows the *arithmetic reasoning* scores of 12 higher-income accountants and 15 lower income accountants. (Two were bashful about reporting their income to us, which is perhaps not surprising. We found varying amounts of this shyness, possibly representing a feeling that we were the latest trick figured out by the Internal Revenue Bureau. It seemed to be most pronounced among the lawyers in our group.) In Slide 3, the more successful accountants are shown in orange above, the less successful in blue below. Note how sharply they differ. The more successful pile up at the "superior" and "above average" levels, while the "average and below" group contributes only what we might call "marginal" accountants. Arithmetic reasoning ability appears important not only for getting into accountancy, but also for getting ahead in it. Remember, the tests were given 13 years ago in 1943; the job measure was obtained ten years later.

Some of the other abilities that were important for getting ahead in accounting are shown in the next three slides. Slide 4 shows *reading comprehension.* This test consists of passages of rather technical reading material, each passage followed by a series of questions to test comprehension. The group as a whole is not as outstanding in this ability as in arithmetic, but the differences between the successful and the unsuccessful groups are about equally marked. Slide 5 shows a test of *reaction speed.* The subject was required to react as quickly as he could to a pattern of lights, pushing one of four toggle switches, the correct switch depending on which pattern of lights was flashed. The successful accountants excelled in this task calling for accurate discrimination and quick response. Slide 6 shows a test of *finger dexterity.* Here, the task was to lift each of a series of square pegs from its hole, one at a time, rotate it through 180 degrees and put it back in its hole. The test was scored for speed of performance. As a whole accountants don't appear to be particularly nimble-fingered, but the successful ones are. At least, they quickly mastered this task of using their hands rapidly and accurately.

Why do we find these last results on finger dexterity and reaction time? Do they make sense, or are they something that just chanced to show up in our 27 cases? I can't tell you, but at the moment, 27 cases is all I have to offer you. More of that later.

By way of contrast, and to show you that not everything predicts success in accountancy, I show you in Slide 7 the *spatial relations* scores of the "successful" and the "marginal" accountants. In this test, the examinee was supplied aerial photos of little segments of terrain, and aerial charts covering larger areas. He was required to spot on the chart the place represented in the photograph. Note that neither group tends to get particularly high scores on this test, and that there are essentially no differences between them.

Enough for accountants. Let us look at two or three other groups. First, I show you a group of 54 engineers —electrical,

mechanical, chemical, mining, civil, industrial and sales. Slide 8 shows the test that best predicts whether a young man will get into engineering —a mathematics test, based largely on problems from high school algebra. But mathematics achievement at age 20 tells rather little about whether a man will get ahead and make money in engineering. In the low math group there are both successful and unsuccessful engineers. Among the tests that we tried, the one that best discriminates levels of success in engineering is the *spatial relations* test —the one that was of no importance for accountants. The results are shown in Slide 9. The engineers as a group were rather high on most tests, and several others, such as reading comprehension and arithmetic reasoning, also predicted later success as represented by reported income.

One group in which I think you may be interested is a small group of owners or managers of firms manufacturing durable goods. Slide 10 shows the *reading comprehension* scores of the eight high and nine low income members of this little group. It takes brains to run a business successfully!

So far we have dealt solely with professional and managerial occupations. Let us look at one or two skilled trades. I have only 12 machinists, but I would like to show you how they do on a test of *two-hand coordination.* In this test, the subject was required to use two lathe-type handles to keep a pointer in contact with an irregularly moving target button. The controls were very much like those of a lathe, though I cannot assert that they were used in just the way a lathe operator would use his. Slide 11 shows the machinists on this test. You might like to see these same machinists on *mathematics achievement.* Slide 12 gives the picture. Slide 13 shows the *finger dexterity* and the *arithmetic reasoning* scores of a group of 25 machine operatives. Slide 14 shows the *mechanical principles* and *mathematics achievement* scores of 19 auto, plane, radio and T.V. mechanics and repairmen. There are many other groups that I could tell you about if time permitted.

Admittedly, the charts I have been showing you are selected

cases. They suggest that *certain* of the 1943 tests gave significant predictions of *certain* of the 1953 jobs and of success in them. Obviously, for any job there are some abilities, often many abilities, that are not of critical importance. I have shown you one or two of these cases. There are also some of the jobs for which *none* of our tests provided any useful prediction. Thus, none of our tests was of any use in describing wholesale salesmen, or in predicting success in wholesale selling. There are also instances in which it is hard to see much of any sense in the test results. For example, our successful insurance salesmen and brokers surpassed their less successful brethren only in tests of motor coordination. Of course, there is the theory that this arises out of the critical importance of golf playing in the career of the insurance salesmen —but if we discount this explanation the result seems rather irrational. Why do we get some of these peculiar results?

This question brings into the open a nasty thought that has probably been lurking in many of your minds while I have been speaking. You have probably been saying to yourself: "Why has this man been wasting my time with his 29 accountants and his 19 repairmen? He starts with half a million men and he ends up with a mere dozen machinists".

This brings us to the next act of our little drama. After two years of peddling our project to most of the major research foundations in the country, just a year ago we finally got from the Grant Foundation funds that will permit us to follow up some 15,000 of these men. The project is in full swing now. Your invitation came a year early for me to give you a report of results, but I would like to tell you what we are doing, what the problems are in a project of this sort, and what we hope to be able to offer the psychological fraternity in a year or so.

The first problem we faced was finding the men for whom we had test scores. Our starting point was the names, army serial numbers and test scores of some 17,000 men. These men were approximately a 20% sample of all men tested by the Air

Force on the Aviation Cadet battery between July and December 1943.

Fortunately, at this point we were able to enlist the cooperation of the Veterans Administration. The names were checked in their locator file and some record was found for all but eight. If the man was still living, the VA undertook to provide an address for him —either the one at which was listed on his GI insurance policy or the one from which he had filed a VA claim. Fortunately, there are two address sources here for most men, so that when one address is out-of-date (as we have found it to be in about a third of the cases) we can try the other.

Men whom we are unable to locate through the VA we shall seek at the Army Records Center. This is a fabulous place, in which the personnel records of some 20 million men who have served in the United States Army or Air Force are kept. And not only are the records kept there; they can actually be found as well. We have used the services of this agency on several projects, and have been uniformly impressed with the speed and completeness with which records are located and made available.

The Army Records Center will give us address at time of separation from the service, and address of next of kin. It will also indicate which men are still on active duty or in the Reserves. For these, we can get addresses from the active military personnel files. After tht "Thorndike Detective Agency" has exhausted these clues, the number of Missing Persons will be quite small, I believe. I would guess that we will have an up-to-date address for over 90 percent of our original group. If time and our resources permit, we can try through postal tracers and the various veteran's organizations to get some line on the remainder. But since we are not really running a Missing Persons Bureau, we will probably be content to accept some loss, trusting that it will not seriously bias our results.

The second problem is to elicit response from the men. So far, we have relied upon questionnaires. We have been at great pains to keep our questionnaire brief —one side of one page.

We have appealed to each man on the grounds of helping his fellow man. We have used three letters and a postcard, and by this combination have gotten responses from over 70 percent of the individuals whom we have reached with our letters. (I estimate that our questionnaire has been actually delivered to about 10,000 men, and we have now gotten back some 7,300 replies. A mailing to some 4,000 at new addresses should bring our total of returns over 10,000.)

For the 25 or 30 percent who will not reply to a questionnaire, more heroic measures are necessary. We are currently embarking upon these. The resource that we are enlisting is one that would not commonly be thought of in psychological research. This is the Retail Credit Company —a credit interviewing concern with branches all over the continental United States. Through this agency, it is possible to reach a person by phone or personal interview at almost any location on the mainland. The interviewers are not psychologists, but our questions are simple and factual. Experience in the pilot study indicates that Retail Credit's interviewers can get answers for us —and at a tiny fraction of what it would cost us to do the work ourselves. If our budget will stretch to cover it, we expect that we can reach most of those who have failed to return our questionnaire in this way. Again, it is my belief that we will be able to get the basic facts that we are seeking for 85 or 90 percent of those for whom we find current addresses.

What can we use as criterion measures against which to validate our tests? For the present we are being content with simple and crude measures. Conceivably we may seek more detailed and more precise information later. Most simply, we have the fact that the man entered and is making a living at an occupation. He *is* a doctor, machinist, or postal clerk, even though he may not be a very good one. We can first see how doctors, for example, differ from the whole group with which we started. We can require further that he must have persisted in the occupation for some minimum period of time —for example, two years. For degree of success we are currently limited to three items that we have asked the man to report to us —his income,

his liking for his job, and his own impression of how good he is at it. As I showed you in the case of the pilot study, we have some indication that these are predictable, at least for certain jobs. They are certainly somewhat unreliable, but are probably not biased, so far as their relation to test score is concerned.

With the data in hand, the final problems relate to how we shall analyze them and how organize the results for potential consumers. The basic pattern of analysis seems fairly straightforward. We shall need the correlation of each of our tests with group membership, and within each group we shall need the correlation with each of our criteria of success. We shall need the correlations among our tests, and we shall need to determine regression weights to find out how much independent contribution each test makes to predicting success in a given job.

For any one job, we shall probably pick the three or four tests that combine to give the best prediction of success on that job, and prepare expectancy tables relating test score to probability of entering and of succeeding in that job. These tables will present essentially the same sort of information as that in the slides that I showed you, only in numbers instead of dots, and for enough cases to make the results reasonably stable and dependable. There should be 30 or 40 job categories for which we will be able to prepare dependable expectancy tables. These will show the counselor or personnel worker what the probabilities are, at any given test score level, that a boy could get into that job and that he could succeed in it once having entered it. I hope that we can express the findings sufficiently simply and clearly so that every psychologist working in the counseling field will be able to understand and use them. In my more exuberant and optimistic moments, I even dream of their being used by the working high school counselor —but my friends and associates assure me that this is the wildest of fantasies.

Our study is far from perfect. We recognize a number of limitations some of which I have hinted at as I have gone along. The most serious are these:

1. We have a select group of individuals, screened by both

a preliminary examination and by their own choice in applying for Aviation Cadet training. The lower ranges of ability are poorly represented, and consequently we have a limited and biased sample of workers in many occupations. We can set no minimum level of verbal ability for truck drivers from our data (if such a minimum exists) because anyone with verbal ability too limited to become a truck driver would have been unable to pass the screening test and get into our group.

2. At best, we will have incomplete returns, and we expect that those returns are biased. We found out about our forger from official records —not from his questionnaire response. The criminal, the mentally disturbed, and the cheerfully well-adjusted bums will rarely return their questionnaires and are unlikely to have their fair representation in our results.

3. Our measures of job success are crude in the extreme. We may fail to predict success because we have failed to measure it. For this reason, generally negative results might lack conviction. But positive results will be all the more impressive, and if we can predict in some jobs, our failure in others will gain in meaning by the contrast.

4. Our Air Force tests correspond only in part with those that the counselor is accustomed to use. In some cases, there will be a problem of translation in giving meaning to the Air Force measures.

With all these limitations, however, analyses based on an actual follow-up after 12 years of 15,000 men tested with some 20 measures of various abilities is something new under the psychological sun. The study should provide a body of data that will strengthen enormously the foundation of fact on which the counselor, the personnel psychologist, the test maker and publisher can build their structures of psychometric practice.