DOCUMENT RESUME

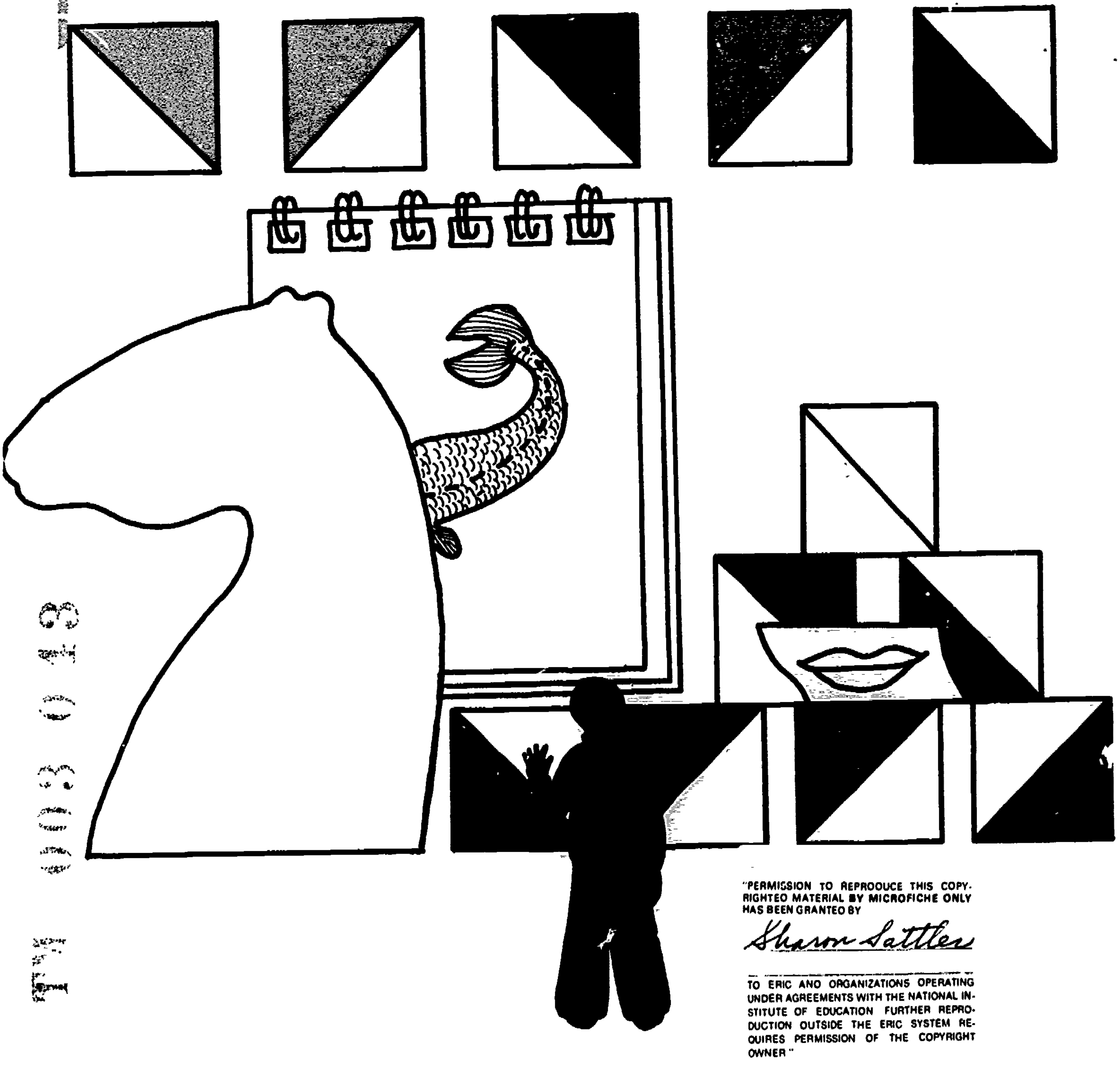
ED 080 538 TM 003 043

AUTHOR	Whaley, Donald L.
TITLE	Psychological Testing and the Philosophy of Measurement.
INSTITUTION	Behaviordelia, Inc., Kalamazoo, Mich.
PUB DATE	73
NOTE	62p.
AVAILABLE FROM	Behaviordelia, Inc., P.O. Box 1044, Kalamazoo, Michigan 49005 (no price quoted)
EDRS PRICE	MF-$0.65 HC Not Available from EDRS.
DESCRIPTORS	*Measurement; *Psychological Testing; Psychological Tests; Sampling; Statistical Analysis; Statistics; Test Construction; Test Reliability; Test Results; Test Validity; *Textbooks

ABSTRACT

An introductory textbook on psychological tests and measurements is presented in paper back booklet form. The style is informal and humorous, and the book is intended to appeal to the contemporary student. Ten chapters constitute the text: (1) On Measurement and Existence; (2) A Brief, Imprecise History of Psychological Testing; (3) The Creation of Differences; (4) Psychological Tests: A Definition; (5) Test Results and Numbers; (6) Descriptive Statistics; (7) Inference Samples, and the Normal Curve; (8) Validity; (9) Reliability; and (10) Constructing a Test. (KM)

Psychological Testing and the Philosophy of Measurement



Psychological Testing and the Philosophy of Measurement

Donald L. Whaley

The staff of Behaviordelia involved in the production of this book were:

PRODUCTION MANAGER:
Daniel Reese

EDITORS:
Helga Ludlow
Robert Ludlow
Lyle Grant
Elizabeth Wolf
Richard W. Malott

COMPOSITION:
Susan Wiltse
Leslie Winnick

LAYOUT:
Diane Hatfield

COVER DESIGN:
Patricia Hartlep

BUSINESS MANAGER:
Sharon Sattler

Publisher: *Behaviordelia, Inc. — P O Box 1044 — Kalamazoo, Michigan 49005 — Phone: 616-382-5611.*

Printer: *Edwards Brothers, Inc. — 2500 South State Street — Ann Arbor, Michigan 48104.*

TABLE OF CONTENTS

INTRODUCTION

Yes, friends, they said it couldn't be done! They said we couldn't write an introductory textbook on psychological tests and measurements that would teach effectively and still hold the interest of the majority of students.

"It's a contradiction in terms!"

"Incredible arrogance! What makes you think you can succeed where thousands have failed miserably?"

"Bah, humbug!"

"You no good, conceited %*&(+@!"

"Man, you gotta be kidding!"

These sentiments, randomly selected from hundreds of equally scornful predictions, would have discouraged the typical, profit-hungry publisher. News of this venture had other ominous repercussions: stockholders panicked; the bank canceled our credit; educators scoffed; colleagues avoided us; students snickered derisively in our classes, the mirth was unconcealed.

Undaunted by these reactions, and the great weight of empirical evidence notwithstanding, we held tenaciously to our theory that testing, measurement and statistical inference could be made relevant – even stimulating – to contemporary students.

Have we succeeded, you may ask, in making this traditionally intractable subject matter both lucid and lively without perverting the principles and methods beyond recognition? Is the resulting material worth teaching at all, assuming that it has been made palatable?

It is our honest conviction that our treatment of these basic principles is suitably rigorous for introductory courses. The concepts have not been watered down; rather, we have exercised a careful selection of subject matter, aiming not to provide an exhaustive survey, but to lay a conceptual foundation for more advanced work. To this end, we have devoted more than the usual amount of space to the development of major concepts. In place of the usual cursory philosophical and historical survey, we have chosen to emphasize these two aspects (philosophy and history), hoping that students will thereby feel more comfortable with the technical details of measurement and statistics.

We don't guarantee success, but we feel that the effort was long overdue. Naturally, we have also retained Behaviordelia's informal, humorous style, and we have ruthlessly tried to excise pedantry wherever it reared its ugly head. These two features should help to reduce student apprehension and resistance.

Try this text in your courses, and send us your reactions and suggestions, so that we can incorporate them in future editions, thereby enhancing the enjoyment and effectiveness of the material.

CHAPTER 1

On Measurement & Existence

Since this book will deal with psychological tests, it is appropriate that the reader should begin by performing a psychological test on himself. First, a pair of sturdy steel pliers should be held in the preferred hand. Next, the jaws of the pliers should be affixed just below the bridge of the nose. Now, the handles should be squeezed with great force. . . Eeyaha! There, the test is complete; the person screaming in agony is you.

Such results should prove conclusively to you that you do exist, but if you are still skeptical, you might try stepping in front of an onrushing Mack truck. The author hopes, if only for the sake of further exposition, that this won't be necessary.

Undoubtedly, many readers will feel that the preceding test should not be an entering requirement for an introductory course in the basic principles of psychological tests and measurements. Certainly something must exist before it can be measured — existence must necessarily precede measurement! But, of course, if it were all that simple, there would be little point in pursuing the subject any further.

REALITY OR ILLUSION

It may well be that most of our problems would vanish if we could only agree on what exists and what does not. Unfortunately, there are a number of obstacles which have prevented us from arriving at a consensus on the question of existence. For example, under certain environmental or physiological conditions, almost all of us will attest to the occurrence of phenomena which experience tells us are impossible. These instances are called *illusions* or *hallucinations*, and they are by no means new to human experience. Philosophers and scientists have puzzled over them for centuries.

One philosopher struggled valiantly with the issue of reality and illusion, and, for awhile at least, lost himself in a tangle of premises, suppositions, and syllogisms. His name was Rene Descartes.[1]

It occurred to Descartes that such events as dreams and illusions, which were later judged to be unreal, were, at the times they occurred, every bit as real as any reality previously experienced. How then, he reasoned, can we be assured that there is indeed a real world? Perhaps the world and all in it were illusion. The possibility had occurred to others before Descartes, and I'm afraid it is still a topic for speculation. Although Descartes failed to reach a solution that was historically satisfying, his attempt to do so is nonetheless well worth recounting.

Through the use of logic and reason Descartes began to search for the proof which would establish the existence of a "real" world. He began by questioning the existence of objects and events which had previously been taken for granted. Soon Rene began to doubt everything. He doubted trees, lakes, streams, fish, and animals; he even doubted his pet frog and pocket knife. And yes, finally he even doubted himself! "Do I exist?" he asked again and again of friends, family, and even passers-by. "Of course, you ninny," they would say. Some of them even gave him a clout on the ear to convince him that he was around. But Rene couldn't make anything of it because he wasn't sure *they* existed!

Now you shouldn't be left with the impression that he doubted everything all of the time. Things would come and go. His ears would flicker and fade, or his fingers and toes would vanish unpredictably. One auspicious morning he was questioning the existence of his navel, which had been waxing and waning for hours. Suddenly, he was rescued from his dilemma by one of his renowned flashes of genius: "Voila," he said, "it seems like the only thing that I don't doubt is my thinking," and he knew at once that he had the answer. For if thinking occurred, someone had to do it; and it logically followed that he had to exist.

"Of course," he said, pounding his bewildered frog with his fist, "Cogito ergo sum! Cogito ergo, sum!" Rene felt a lot better after that, and pretty soon he was once again able to have faith in all those things he had previously doubted. One by one, he regained his ears, his pocket knife, his pet frog, the trees, the streams, his family, and finally, not without reluctance, his local draft board.

However, before long Rene's friends and companions got fed up with him. They said, "Cogito ergo sum, cogito ergo sum! Is that all you got to say for yourself? Why don't you get rid of that filthy frog and get a job?" His feelings hurt, Rene left the city and returned home to his mother. She had been right all along. Thinking was dangerous.

Obviously Descartes' efforts to establish the existence of reality have been improved by several centuries of modern science and technology. Modern man can provide reams of computerized data which conclusively "prove" his own existence. Indeed, he has the data, but none of it even begins to establish scientifically that things are other than fantasy or illusion. The question, "Do we exist?", is still open, and we will return to it time and again to find that it has still not been resolved.

AWARENESS OF EXISTENCE

In the meantime, let us consider a few historical facts with which most of our friends and colleagues (whether they are phantoms or not) can agree. In contemporary society, things exist in our conscious awareness which did not exist for people living in earlier societies. Atoms, microbes. the planet Pluto, the Oedipus Complex, DNA, atomic radiation, vitamins, and microspheres are just a few. How did these objects and concepts come to exist in our collecitve awareness?

Undoubtedly you will say that they were discovered. In some cases, you may even supply the name of the discoverer and the date of discovery. Individuals first discovered or noticed them and then made their existence known to others. Let us examine a few less famous "discoveries".

JAMES THURBER DISCOVERS THE BEASTS OF THE SLIDES

In one of his autobiographical works, the late American humorist, James Thurber, tells about some of his problems concerning existence.[2] Thurber had always suffered from poor vision, being forced to wear thick-lensed glasses for most of his life. Even with this corrective device, his vision was inadequate for many purposes. His problems with existence began in a biology course which required experience with a microscope. Students were expected to peer through the microscope, identify the organisms on the slide preparation, and make drawings of them on sketch pads.

Thurber's first difficulty arose when he mistook the pencil sharpener for the microscope. (Possibly he was trying to sharpen his eyes.) A concerned fellow student led him away from the pencil sharpener and positioned his head properly above the microscope. Thurber tried to fake it for some time by emitting appropriate "oohs" and "aahs" which he hoped would be taken as evidence of his entry into the fantastic zoological kingdom of the slide.

But this did not satisfy the lab instructor, who was not going to let a case of simple blindness stand in the way of important instruction. He stationed himself at Thurber's elbow, insisting he see the flagellae at the center, the cytoplasmic vacuole at 3 o'clock, the chromoplasts at noon, etc. Coerced into performance, Thurber reluctantly came up with a group of sketches depicting organisms more bizarre than any ever captured on a slide. The beasts he finally "saw" were so terrifying that the lab instructor fled biology altogether, developing at the same time an acute zoophobia.

The case of James Thurber may seem irrelevant. It is, one may declare, a simple case of faking the data. In order to obtain acceptable grades, Thurber invented the creatures he drew on his sketch pad; but a careful reading of his account leaves some doubt about this explanation. There is the distinct impression that at one point Thurber believed he did see some of the organisms he depicted in his sketches, and so it is plausible to suppose that at that moment those organisms really existed for Thurber. If you had been in Thurber's place, would they have existed for you?

THE FABRIC OF FANTASY

For our next episode, in which things or events which formerly did not "exist" were brought into awareness, we must travel to a different time and place. A young physician was nervously preparing to meet his first patient. Financial considerations had forced him to abandon a promising research career and open a private practice; but he could not also relinquish his beloved theories, which had for so long been the main object of his prodigious intellectual efforts.

The patient's name was Frau Schroeder. Her husband was a wealthy Vienna banker of long-established reputation. She had heard about the new physician from a friend, who had in turn received a recommendation through another physician in the city. Frau Schroeder had been to at least a dozen other physicians in the past two years, none of whom had been able to help her in the least. Headaches, dizziness, and periods of time in which she could not hear were her chief complaints.

"Ah, Frau Schroeder," Dr. Freud said, "Please. . ." He pointed toward a couch and indicated to the plumpish matron that she should lie down.

"Was ist das?" questioned Frau Schroeder. Her face flushed. The doctor was so young, a fact not well concealed by his fashionable beard. He was not at all the grandfatherly figure she had come to expect in physicians.

"Don't be alarmed," Freud assured her with a comforting smile. Frau Schroeder self-consciously positioned herself on the couch.

"Now," the young physician said, "I want you to relax and tell me the first thing that comes into your mind."

"But Herr Doctor," Frau Schroeder said, "I am so confused."

"You complicate things unnecessarily, Frau Schroeder. You must first relax, and then you must tell me whatever comes to mind. Remember, relax!"

Frau Schroeder smoothed her dress a half-dozen times, checking to see that the hem was not indiscreetly raised above the ankle. Dr. Freud waited patiently as furrows appeared on the worried matron's face. Finally, Frau Schroeder relaxed somewhat, and she even managed a slight smile. "Let us see," she said. "Ah, yes, for tonight Hilda is to prepare the fine sausage Herman brought from Frankfurt. Along with them, the pastries should set well for dessert. And then, for Sunday dinner when Herr Munder comes, I think a roast . . ."

"Nein, nein," Dr. Freud interrupted. "Let us not hear about your dinner now. Perhaps you could think of something else."

Eventually, after months of effort with Frau Schroeder, Dr. Freud uncovered some startling and remarkable information through the use of a new method he called "free association". Frau Schroeder revealed a childhood episode in which her father had made improper sexual advances toward her. A remarkable breakthrough, Dr. Freud believed. Now he felt more certain than ever that problems like Frau Schroeder's dizziness and loss of hearing must be due to such early traumas.

Soon Freud had collected data on 18 patients with symptoms similar to Frau Schroeder's. In all of these cases, free association led finally to the report of an early incident involving incestuous sexual offenses.

Freud now believed he had proven his point in the bitter dispute which had ended his collaboration with Josef Breuer.[3] His data were presented to the Society of Psychiatry and Neurology in Vienna in 1896.

Returning to his work, Freud continued to find case after case in which patients reported lurid accounts of sexual experiences in childhood — always instances in which the patient had been victimized by unscrupulous or demented adults.

As the data accumulated, Freud grew uneasy. Something was wrong. Finally, he faced a bitter reality. Could it be that incest and perversion occurred so commonly? After all, his patients were from fine families. Their parents were respected members of society, the bulwark of religious and moral leadership in the community. Freud concluded that it could not be so. The accounts his patients had related to him were fictitious. They had never occurred. But Freud was equally sure his patients had not willfully lied. No, they believed in their own accounts; to them they were real; they had happened; they did exist!

Let us put Freud's dilemma into perspective. Here were his patients, genuinely tormented by their problems, faithfully attending weekly therapy sessions which were never anything but difficult for them. Furthermore, the treatment was quite expensive. Yet, in the face of these apparent hardships, patients wasted much of their time manufacturing fantastic lies. These lies often cast parents or close relatives in the most despicable role — that of perpetrating sexual crimes on a defenseless child. Finally, such accounts — obviously irrational and, in some cases, absolutely impossible — were accepted by the patient as an accurate chronicle of his past.

Freud's attempt to reconcile these facts was no less than fantastic. He concluded that the episodes the patients remembered were not real happenings, but fantasies which occurred when the sexual desires and wishes *of the patient* were not reciprocated by the parent. Yes, even as infants, patients had desired sexual relations with the parent of the opposite sex. These desires were satisfied in the world of private fantasy. These fantasies were what the patients remembered and recounted as real in free association.

The conclusion reached by Freud has

subsequently been severely criticized and hotly disputed. It is not, however, our purpose here to discuss psychoanalytic theory. There is some similarity, however, in the way in which Freud's patients became aware of these fantasies they believed to be real and the manner in which James Thurber became aware of the beasts of the slides. The lab instructor at his elbow insisted on a reality Thurber attempted to capture on his sketch pad. Could it be that Dr. Freud, hovering over his patients, had in a more subtle fashion insisted on the reality they ultimately "remembered"?

Those reading this text who have the unsavory habit of trying to decide just where it is the author is attempting to lead them, may have reached a tentative conclusion.

"So that is his game!" they may say. "He is trying to get us to believe that things come to exist for us merely because authority figures insist we 'see' them! Well, bosh, fol-de-rol, and poppycock!"

Before you jump to such a premature conclusion (you notice the conclusion is termed "premature" and not "incorrect"), let us examine another discovery.

THE IMPOSSIBLE POSITRON

In physics, positively charged particles, or protons, are opposed in a nuclear configuration by electrons, which hold a negative charge. It was theoretically impossible for a proton to have a negative charge or an electron to have a positive charge. Such a thing could not exist. At this time the measurements or observations in nuclear research were made through what was essentially a photographic process. The existence of particles was verified by repeated appearances of similar marks of exposure on the photographic plate. Proton traces were much larger than electron traces. In addition, one could tell the charge on the particle by the direction of the mark on the plate. Particles with a positive charge left a trace angling in one direction, while negatively charged particles angled in the opposite direction.

For many years researchers had encountered a common phenomenon. Traces of relatively small size repeatedly appeared on the plates. These traces were similar to those left by electrons, but the angle of the marks was always in the positive direction. When scientists saw these plates, they immediately disregarded the traces because of the contradictory configuration they exhibited. "It must be our technicians screwing up again," they said to each other; and shaking their heads, they lamented their inability to secure competent help.

One physicist, Dr. Anderson, initially reacted in the same manner.[4] "It must be our technicians screwing up again," he said to his colleagues upon encountering the curious traces. Later, when the plates were shown to the technician, he shrugged his shoulders. "That's the way it is," the technician said. "I didn't make any mistakes."

"That's preposterous," Dr. Anderson replied. "If what you say is true, it would mean that there must be an electron with a *positive* charge, and there's nothing in all of atomic theory to substantiate such a hypothesis!"

"Yeah," the technician said, shrugging his shoulders again. "That's your problem, Bub."

The technician's insistence on the reliability of his techniques led Anderson to examine nuclear theory and to alter it in such a manner as to allow for this new possibility. In revising the theory, and in further checking it out with other experiments, the positive electron, or positron, came into *formal* existence. With it, other possibilities became immediately apparent, thus paving the way for other important discoveries in nuclear physics.

This discovery is different from the two others we have discussed. The positron had existed (insofar as any particle exists on a photographic plate) for many years. Yet, its existence was categorically disregarded and denied. It came into existence for Dr. Anderson when he considered that the traces on the plate were not just errors due to sloppy technical procedures. When Dr. Anderson modified the theory to accomodate the existence of the tracings, the possibility of a positively charged electron became apparent to other scientists. Subsequent research confirmed its existence.

Certainly the positron did not exist because of the intimidation of an overzealous lab instructor, nor did it emerge from the shadowy past through the gentle but insistent beckoning of a Dr. Freud. No one wanted the positron to exist, least of all the scientists who saw it on the photographic plate as a stubborn blemish they wished they could erase.

In spite of the apparent lack of similarity among the three "discoveries" we have just discussed, it is the premise of the author that there are *common features* which account for the emergence of these events and entities into the conscious awareness of the individuals involved. Furthermore, a "heroic" effort will be made throughout this book to demonstrate the common procedures which gave James Thurber his beasts, Freud's patients their memories, and physics the positron. These procedures are not substantially different from those which make it apparent to us that one piece of steel is longer than another when a ruler is applied, or that one stone is heavier than another when a scale is employed. Briefly stated, the principles of human experience which make objects or events exist for us in the first place are effectively the same as those which operate when a measuring procedure brings forth the judgment that objects are bigger, smaller, heavier, longer, or otherwise "different" from each other.

BACK TO REALITY — RENE RIDES AGAIN

It would be remarkable at this juncture if the reader were willing to calmly accept the notion that Thurber's beasts and inches of steel or pounds of flesh are produced by similar procedures.

"Inches of steel and pounds of flesh are real," the reader will maintain. "Thurber's beasts were lies told to satisfy the impossible demands of an academic martinet. They definitely were not real!"

So once again we must return to Rene Descartes nestled safely in his reality of "Cogito ergo sum". What is real and what is not? Descartes seemed to be satisfied with his solution to the reality — illusion problem. For him the dilemma was resolved by the self-observation that he was "thinking". From this point, the existence of a thinker (Descartes himself) and the rest of the world followed logically.

Descartes' proof did not satisfy subsequent philosophers and scientists. Science renounced Descartes' "deductive" or "rational" method of inquiry, wherein issues were attacked with formal logic and reasoning. Contemporary science insists on the "empirical" approach, in which exhaustive observations of a *public* nature are the only acceptable proof or data. Descartes' proof may have been adequate for him, today's scientist suggests, but since the "thinking" process which ultimately convinced him of reality cannot be observed by others, it is not admissible evidence. Lamentably, however, modern science does not, in the words of an American political figure, "offer reasonable alternatives". In fact, most modern scientists agree that there is no way by which empirical methods can be employed to prove that the environment or any part of it actually exists, even though those same methods have enabled us to achieve very effective and reliable control over a substantial portion of that environment.

BERKELEY'S BOX

George Berkeley, a philosopher who lived and wrote in the mid-1700's, summed up the difficulty in proving an independent reality through objective and public observation.[5] It was his contention that we can never know if what we see, feel, touch, or otherwise experience is as it appears, or for that matter, exists at all. What we experience is not "direct", but is "filtered" through our own sensory system before it exists in our conscious awareness. Indeed, even if others apparently experience the same objects and events, they, like ourselves, are insulated by their sensory apparatus. Their agreement in no way proves things are really that way or that they have an independent existence. Agreement proves only that we are in the same "box".·We are all trapped in a "box" which keeps us from seeing the true reality or discovering that none is there. Since we can never escape the confines of the "box", we can never know in any objective manner that which truly exists.

A WORKING STRATEGY

In the following chapters an attempt will be made to:

1. Introduce principles of psychological testing and measurement.
2. Relate those principles to measurement techniques of other sciences and disciplines and to the world of practical affairs.
3. Relate measurement to principles of behavioral psychology which account for human experience, including that experience not directly concerned with measurement.

Several different issues must be discussed, and concepts not generally included in books devoted to psychological testing must be introduced in attempting to achieve the aims mentioned above.

To begin with, a case will be developed with the intention of persuading the reader that objects, events, concepts, and other phenomena come to exist in the conscious awareness of *any specific individual* because a common set of behavioral principles is operating. This emergence into existence occurs independently of whether it also "exists" or would be judged "real" by others. However, the agreement of others may play an important role in the continued "existence" of the thing or event for the individual.

Secondly, if these common behavioral principles operate for groups of individuals, the particular entity which previously existed for only one individual will emerge fo. the group. In a practical sense, "reality" is defined by agreement among relatively large numbers of people. In addition, proof of "real" existence is the observation that the new entity or phenomenon is associated with or related to other occurrences or entities which are already accepted as "real", or *bona fide.*

These two points are relevant to the "discoveries" discussed earlier. The "Beasts of the Slides" existed for Thurber when the lab instructor insisted upon a reality under the microscope which Thurber was expected to perceive. The beasts were subsequently judged to be imaginary when they did not resemble sketches made by other students, and when others rejected the possibility of their existence.

Incestuous incidents came to exist in the personal histories of Freud's patients when he subjected them to procedures of "free association". They existed as "real" even for Freud until their existence was judged inconsistent with other "real" situations in the typical homes of his clientele and general culture in which he lived.

Marks on photographic plates existed for scientists and technicians alike. The positron did not "exist", however, until those particular marks were judged by Dr. Anderson to be highly persistent under well-controlled laboratory conditions and were therefore rejected as being attributable to imprecise or careless procedures. Dr. Anderson conducted more experiments, suggested changes in theory which would allow the possibility of a positive electron, and published experimental results which established the positron's existence for other scientists.

This analysis should lead the reader to the realization that the conditions which make an entity "exist" at a given instant in time, and the conditions under which this existence will be judged false or real, are conceptually separate.

Finally, since in a practical sense reality comes to be defined by agreement among individuals on what is real, and by consistency between what already is judged real and a new phenomena, we are in no position to vouch for an ultimate reality.

We are tightly confined, it would appear, in Berkeley's Box. Unless, like Descartes, we can accept an ultimate reality because of our private belief, we are destined to be forever held captive.

Hopefully, you will realize that existence is not a cut-and-dried issue. Even now, you should appreciate Rene Descartes' dilemma, the monstrous microbes that James Thurber brought into existence for the sake of his lab instructor, and the "impossible" positive electron which existed only because a stubborn lab technician insisted on his own skill.

FOOTNOTES

[1] Vrooman, Jack. *Rene Descartes. A Biography.* Putnam: N.Y., 1970.

[2] Thurber, James. *A Thurber Carnival.* Holt: N.Y., 1961.

[3] Cohen, Jozef. *Personality Dynamics.* Rand McNally: Chicago, 1969.

[4] Hanson, N.R. *The Concept of the Positron.* Cambridge: London, 1963.

[5] Berkeley, George. *Principles of Human Knowledge.* 1710.

CHAPTER 2

~~A~~ Brief, Imprecise History of Psychological Testing

PRE-SCIENTIFIC TESTS

Long before yardsticks or rulers were invented, human beings found it advantageous to discern differences among themselves, and it was natural that priority should initially be accorded to differences in *physical* characteristics. There were obvious advantages, for instance, in being able to distinguish a friend from an enemy or a receptive female from a hungry baboon.

It is from these primitive beginnings that the first selection procedures evolved. The observation that differences in appearance, strength, or agility were related to *later performance* as a hunter, cavemate, or galley slave, led inevitably to personnel selection.

The transition from simply observing appearance, feeling muscle, or checking the sturdiness of bones and teeth to the construction of crude tests of strength and agility was probably a gradual one in the history of personnel selection. Such tests, however, were to become elaborate and highly structured thousands of years before psychology existed as a discipline. One writer has offered a detailed account of what was essentially a system of civil service examinations used in the Chinese Empire for 3,000 years or more.[1]

INDIVIDUAL DIFFERENCES: FRANCIS GALTON'S CONTRIBUTION

It was not until the late 19th Century that the connection between certain kinds of individual differences and differences in performance became a *formal* concept. Remarkably, the concept was called "individual differences".

The formalization of the doctrine of individual differences was the work of one of the real superstars of scientific inquiry. His name was Francis Galton.[2] Galton had read every available book on an incredible variety of subjects, and had traveled extensively. As a result, people flocked to him to hear his opinions on numerous topics. He was an all-around scholar, the prototype of the English gentleman-scientist. It was inevitable that his accomplishments would earn him royal recognition, and thus, today, Galton is known as Sir Francis Galton.

Galton, who was a child prodigy, was intensely interested in the science of breeding when he was a boy. He was introduced to this topic by his famous cousin, Charles Darwin, whose painstaking scientific research and influential writings provided the foundations for evolutionary theory as it is currently taught. Galton wrote numerous papers in behalf of Darwin's theories, often in rebuttal to strident religious attacks against his cousin's work.

Ironically, Galton's side interests contributed most directly to the area of psychological testing and measurement. He was quite a horse-trader and racing fan, and not above betting on the horses. One source, albeit not too reliable, has even given him credit for the invention of parimutual betting. In any event, we know for certain that Francis Galton collected a great deal of information on the breeding practices of horses and other livestock. Concomitantly, he developed mathematical formulations which led to the description and explanation of the *normal curve.* The normal curve has been singularly important in instances where statistics have been applied to human performance or to human physical and psychological characteristics. It is, indeed, still the basis for statistical analysis and application in the behavioral sciences.

EUGENICS

Although Francis Galton resembled a butterfly in his work, flitting here and there as his curiosity dictated, he had as his primary objective the lofty motive of the improvement of the human species. It was easy for him to see that other Englishmen were not so gifted as he, and it was equally apparent to him that "backward" peoples suffered dreadfully from deprivation of the enlightenment that the English enjoyed. He sought to improve the lot of all Englishmen, and perhaps, the entire human race. To accomplish this, it would only be necessary to improve the stock through selective breeding, or eugenics, as it is called.

Selective breeding had, after all, been practiced with lower animals for centuries. For human beings, of course, it presented unique problems. It would be necessary to establish criteria for the selection of individuals to whom the task of propagating the species would fall. Where horses were concerned, speed, strength, and stamina were obviously important criteria. It was not clear, however, that if the same standards were applied to human breeding practices, a race of supermen would be forthcoming. Another possible outcome of such practices would be a race of strong, speedy, and tireless, nitwits.

GALTON'S TESTS

Galton eventually reached a stage in his work where he began to develop instruments and techniques to measure human *abilities.* He invented dozens of new pieces of apparatus and tests which ranged from a task as simple as striking a piece of metal with a hammer to complex tasks requiring delicate movements of complicated paraphernalia.

Galton's tests were placed into a more or less standard sequence, and were then given to multitudes of Englishmen. It was his dream that every citizen in the British Isles would take his tests, with the results recorded for posterity. It was a dream that was not to come to fruition.

Galton's service to the progress of testing was to give it scientific status, both in theory and method. His most important contributions were the discovery of the normal curve and the statistical approach to testing which ensued, as well as the development of numerous devices and other scientific instruments, many of which are still in use today. But for all his science and all his inventive genius and insight, Francis Galton was impractical. For a man who appeared to know exactly what he was searching for, his aristocratic propensities often directed his immense creative talent into very unproductive enterprises. We will have more to say about this remarkable historical figure in a latcr section.

THE FRENCH TAKE THE BALL

While Galton was suffering from his chronic case of impracticality, the French were unpretentiously hitching the psychological testing movement to the plow. French physicians, such as Esquirol[3] and Sequin[4], studied and worked with individuals who today would be classified as mentally retarded. The efforts of these Frenchmen were essentially pragmatic: They were concerned first with specifying differences between handicapped individuals and others in the population, and then relating those differences to future performance in socially significant areas like education and vocational training

BINET HAS HIS DAY

At the turn of the century, another Frenchman named Alfred Binet was attempting to develop tests of intelligence. First, Binet worked with a friend called Henri. Binet and Henri tried many approaches and procedures in order to discover test items of performance which would relate to their conception of intelligence. Such things as handwriting analysis, palmistry, and astrological calculations were tried. Eventually, Binet got rid of Henri, mostly because no one could remember his last name. "Henri who?", they would all ask.[5]

In collaboration with a new partner, Theodore Simon, Binet proceeded under a contract from the French Minister of Public Instruction to devise methods to educate sub-normal children in the public schools. This assignment ultimately led to the development of the Binet-Simon scale[6], which was published in 1905. Binet soon became the authority on mental testing, and today our most cherished intelligence test differs but slightly in content from the original test developed by Binet and Simon. While Binet became an historical great (undoubtedly making a bundle in the process), Simon dropped mysteriously from the scene shortly after publication of the first test. There ensued quite a bit of speculation about Simon's role in the development of the scale. Some authorities feel that Simon's interest in intelligence resulted from over-compensating for his own real or imagined inadequacies. Some have gone so far as to suggest that he was the original "Simple Simon", which seems highly improbable, however.

CATTELL AND AMERICAN TESTING

Although testing originated in Europe, it was soon adopted in America, quickly becoming a naturalized citizen. Americans displayed an immediate affinity for the psychological test because it seemed to have the potential to become a useful practical tool. Regardless of its origins, the testing movement was uniquely an American episode; it was and is a slam-bang, smack-dab, star-spangled American venture. Like the hot dog, it was taken over lock, stock, and mustard.

Concurrently, the scientific and theoretical underpinnings of the testing movement, which had been established by Galton, fell heir in America to James McKeen Cattell.[7] For Galton, the study of individual differences had been one of many side interests; for Cattell, it became a crusade. Cattell relentlessly attacked and critically wounded the entrenched, dominant psychology of America — that of Titchener at Cornell — which had been imported from Wundt at Leipzig. Cattell's principal objection to Titchener's "structural psychology" was that it failed to respect individual differences. The fact that two individuals behaved differently on the same task was, for Titchener, as it had been for Wundt, an unfortunate condition resulting from experimental error. In the best of all worlds, in the best of all laboratories, on the best of all days, their responses would be identical.

For Cattell these differences were not error; they were not incidental. They were real, and as such were the proper content for psychology. Cattell proved to be a formidable adversary for Titchener. He was brilliant and industrious, having served as Wundt's assistant when studying in Europe, and he understood the opposition perhaps better than it understood itself.

For all his steadfast zeal, Cattell did little personally which made any practical contribution to psychological tests as we know them today. He was far too much the academician. Perhaps his most important achievement was to establish Columbia University as the fountainhead from which issued a succession of formidable figures in American psychology for decades.

ANIMAL INTELLIGENCE

The most notable of Cattell's students was E. L. Thorndike. As a breed of robust people, fresh from co-habitation with the beasts of the wild frontier, Americans had developed a keen interest in animals and pets of all kinds. One of the most controversial aspects of the testing movement to come out of this national heritage was an assessment of animal ingenuity. Thorndike entered into the area of animal testing with his book, *Animal Intelligence*.[8] He whetted a great deal of curiosity about how smart animals really were. Soon the topic of conversation around the barber-shop or local saloon revolved around such noteworthy topics as which was smarter, a walrus or an orangutan? Or could a blindfolded pigeon find his way home in the dark faster than a bat could in the daytime?

Soon a man of sufficient stature came forth to answer some of these pressing questions. His name was Hunter.[9] By an ingenious method called the *time-delay experiment*, Hunter attempted once and for all to settle these bitter disputes which were pitting neighbor against neighbor, brother against brother.

The time-delay experiment worked in the following fashion. The animal was retained at one end of a box. The other end of the box was divided into two sides with doors which normally closed off the front end of the box from the back. Food was placed in one of the side compartments but not in the other, and the animal was allowed to watch this. The animal was restrained for various periods of time; then released and observed. If he made a bee-line to the compartment where the food had been placed, it was assumed that he "remembered". Animals were compared in terms of how long they could be restrained before they forgot. Racoons remember longer than rats, but not as long as monkeys. Tigers couldn't wait for any length of time at all, but then Hunter wasn't keen on making them wait anyhow.

INDIVIDUAL AND GROUP INTELLIGENCE TESTS

The Binet intelligence test was adopted in America and revised by Terman and Merrill in 1937.[10] Later revisions were made by Terman and his associates at Stanford. Today, the Stanford-Binet continues to be the most venerated of intelligence tests.

The Binet was an *individual test* which was expensive to administer and this prohibited it from being applied to large numbers of people. While Henry Ford revolutionized manufacturing with the assembly-line concept, Robert Yerkes developed the *group test*. Now large numbers of individuals could be tested. As the nation became involved in the Great War, thousands of recruits and inductees were subjected to the Army Alpha group intelligence test. National leaders were appalled to find that our doughboys were not only brave and robust, as the Alpha revealed, but also stupid. Results from the Army Beta, a form of the same test developed for illiterates, were no more heartening.

INDUSTRIAL PSYCHOLOGY, FREUD, AND INKBLOTS

At this hour, America was going through a "rags to riches" period, of expanding commercialism. Many had made fortunes overnight. One sure way for the boy to make it big was to be able to sell the stream of new products to a gullible public. But a special breed of person was needed. Soon it was apparent that a young man needed more than a bright smile and well-shined shoes to get along in the world.

Thus, personality testing became an important American product. In these early times, personality was not so much a cluster of behavior patterns to be described and analyzed as it was a type of possession or tool to be used as a means to an end. This practical conceptionalization of personality did much to establish industrial psychology and personnel management; and thus, solidify ongoing symbioses between business and psychological testing.

Then the depression brought hard times to young America. It was a sobering experience for a hearty people. As they looked down at tightened belts, Freud insisted they look yet lower. The unzipped knickers, he insisted, were not the result of simple absent-mindedness. The infantile exhuberance industrial America had exhibited was not what it had appeared to be. Personality tests were developed to tap the machinations of that self which operated below our awareness (and below our belt line) even as we slept in our cradle. The most famous of the projective tests, as they were called, was the Rorschach.[11] The Rorschach consisted of standard ink blots constructed to present the subject with ambiguous visual stimuli.

TEST PARANOIA

Projective tests and the theory they were based on helped to cloak personality testing — and ultimately psychological testing in general — in an envelope of mystery. It was conceivable that in naively relating what one saw in the ink blots, private, confidential, and perhaps incriminating information would unwittingly be revealed. But wasn't this an invasion of privacy? Didn't the Fifth Amendment protect the citizen from such self-incrimination?

In his book, *The Hidden Persuaders*[12], Vance Packard revealed and undoubtedly heightened a paranoia in the American citizenry. Packard alluded to unscrupulous interests who were employing Freudian principles to force the citizenry to reveal potentially damaging information about itself and these unscrupulous interests were using the same principles to whet subconscious appetites of the citizenry for the products they wished to sell. These practices were, according to Packard, dirty pool, since they worked at the subconscious level. The consumer or citizen was a helpless and unaware robot in this insidious scheme.

As a result of Packard's accusations, general misinformation, and the effort on the part of some professional psychologists to reify and keep secret from the public the nature of psychological mea- [illegible] psychologists ~~and~~ their tests [illegible] evoke unnecessary fear in the [illegible]

In future chapters an attempt will be made to dissect the odious "mental test" Hopefully, you will have the fortitude to witness the spectacle.

FOOTNOTES

[1] DuBois, P. H. "A Test-Dominated Society: China 1115 B C. — 1905 A. D." In A. Anastasi (editor), *Testing Problems in Perspective.* Washington: American Council on Education, 1966. pp. 29-36.

[2] Pearson, K. *Life, Letters and Labours of Francis Galton.* I, 1914; II, 1924.

[3] Esquirol, J. E. D. *Des Maladies Mentales Considerees Sous Les Rapports Medical, Hygienique, et Medico-Legal.* Paris: Bailliere, 1838, 2 vols.

[4] Seguin, E. *Idiocy: Its Treatment by the Physiological Method.* Reprinted from original edition of 1866. New York: Bureau of Publications, Teachers College, Columbia University, 1907.

[5] Binet, A. and Henri, V. "La Psychologie Individuelle." *Annee Psychologique.* 1895, 2. pp. 411-463.

[6] Binet, A. and Simon, T. "Methodes Nouvelle Pour le Diagnostic du Niveau Intelectuel des Anormaux." *Annee Psychologique.* 1905, 11. pp. 191-244.

[7] Cattell, J. "Mental Tests and Measurement." *Mind.* 1890, 15. pp. 373-380.

[8] Thorndike, E. L. *Animal Intelligence.* 1911.

[9] Hunter, W. S. "The Symbolic Process." *Psychological Reveiw.* 1924, 31. pp. 478-497.

[10] Terman, L. M. and Merrill, M. A. *Measuring Intelligence.* Boston: Houghton Mifflin, 1937.

[11] Rorschach, H. Translated by P. Lemkau and B. Kronenberg. *Psychodiagnostics: A Diagnostic Test Based on Perception.* Berne: Huber, 1942. 1st German edition, 1921. U. S. distributor: Grune and Strutton.

[12] Packard, Vance. *The Hidden Persuaders.* New York: Vantage, 1961.

CHAPTER 3

The Creation of Differences

PROMISES, PROMISES

In the first chapter, some high-sounding promises were made regarding the objectives of this book. They were to:

1. Introduce principles of psychological tests and measurement.
2. Relate these principles to methods of measurement employed in other sciences, disciplines, and in the practical world.
3. Relate measurement to principles of behavioral psychology which account for human experience, including that experience not concerned with mea rement.

AND A FROG SHALL LEAD US

Gossip is usually unwholesome, unless, as in the following case, it results in important educational elucidation and discovery. This tidbit describes some rather peculiar circumstances in which a frog came to his end. This particular frog belonged to Robert Yerkes,[1] an early psychologist who, as we shall see later, played an important role in the development of psychological tests. Yerkes' rather unwholesome relationship with the frog is only tangentially related to his work in testing. The frog had a wild crush on Dr. Yerkes, although, if we judge on a strict behavioral basis, we must conclude that this fond regard was not reciprocated by Yerkes. While it is not exactly clear how it came about (some say it was a lover's spat), Yerkes took the frog and placed him in a sauce pan containing water which was approximately at room temperature. A gas flame was placed under the pan, slowly increasing the temperature of the water. The frog sat calmly, staring with huge adoring eyes at Dr. Yerkes, while around and within him the temperature rose steadily. Within a few short minutes, it was all over for the frog. He was boiled alive.

During the feast which followed, Yerkes munched on a frog leg and explained to an observer that any frog, love-smitten or not, would jump immediately if placed abruptly into boiling water. In this case, the change of temperature from 72° to 212° had progressed in such a gradual manner that, behaviorally speaking, the temperature difference did not exist for the frog.

In many instances humans behave as Yerkes' frog. It is mainly through the reports of others that we gain knowledge that we are older, fatter, or more stubborn. In the moment-to-moment, day-to-day experience of living with ourselves, changes are lost and differences undetected. Similarly, when changes occur rapidly, they may also not exist for the observer. If this were not so, movies would appear as sequences of discreet slide projections and television as a series of light dots systematically changing in intensity and position on the viewing surface.

By and large, however, the ability of Man to behave differently to different aspects of his environment seems to have been to his advantage throughout the course of evolution. The advantage of behaving differently becomes even greater as new technology allows us to make more extensive and more refined distinctions about the world.

Suppose Yerkes' unfortunate frog had been equipped with one of man's simpler devices, the thermometer, along with the skill to read and interpret it. Occasional, brief glimpses of the thermometer would have allowed the frog to make a casual exit long before the temperature of the water reached the danger level, thus averting the tragedy reported above.

The trick to mastering the environment is in first detecting its different aspects, and then responding accordingly in selective ways — ways which promote the achievement of individual and cultural goals.

Differences in the environment can be analyzed logically. Naturally, the very term "difference" implies some type of *comparison*. Some comparisons are successive. The frog failed to respond to a difference which can be viewed as a successive comparison. A situation at one time differs from the same situation at another time because something has been added or taken away. Such a comparison and the resulting difference defined by it is called *change*.

Change often involves the addition of an entirely new entity — one which emerges in an old situation, one which has not been seen nor experienced before. Suddenly, one becomes aware of a new thing or object which appears as separate and distinct. It is as though you are looking through a high-powered telescope pointed toward a distant planet . . . You see nothing but a blur and what appears to be your eyelash. You begin to adjust the focusing mechanism, then suddenly . . . Bingo! As if from nowhere the distinct form of the planet comes into existence. Again, this kind of difference emerges as a successive comparison between what was experienced before the telescope was placed before the eye and properly adjusted, and that which was experienced afterward.

But the emergence of a new phenomenon does not necessarily entail the telescope or electron microscope. It may take place without such devices, and is, in fact, a common occurrence, as when we re-visit old haunts and find, what are for us, entirely new features. Sometimes new features do not come into existence for us. This can be a cause for extreme consternation. The following episode is a case in point.

WHAT I DID ON MY SUMMER VACATION

It is probably no great secret to you that a great many students must work during summer vacations to earn enough money to continue school in the fall. Let us presume that one summer this fate befalls you. In the city where you live, there is a large manufacturing interest which produces electronic transistors for pocket radios. This summer you apply for and are fortunate enough to get a job. You are told that you will be an "inspector". Your job is to reject bad parts and pass on good ones. On the first day of work, your supervisor, who is anything but talkative, takes you to your station near a conveyor belt. Small electronic parts go by and you peer at them through a large magnifying glass. The supervisor explains that you should take the "skags" and put them in a nearby bin, but should let those which are acceptable continue on the conveyor belt to the next station where they will be packaged. Before you can look up, your supervisor has disappeared, off to do his many busy chores, and you find yourself alone, watching the brightly colored ceramics passing behind your minature viewing lens.

Your job, it would appear, is a relatively simple one. Furthermore, you are being paid $5.00 an hour to engage in this activity. The company has even seen fit to provide you with a comfortable chair and pillow. As you settle down to begin work, you take a good look at the transistors, attempting to determine which ones are faulty and which are not. Horrors! They are all the same color, the same size, and they have identical markings. After two hours of watching the parts go on their way, you become somewhat uneasy and begin to look around for your supervisor. But he is a busy man, and is nowhere to be found.

Meanwhile you look across the aisle and see another person about your age who appears to be doing the same kind of work. You discern from his sweatshirt that he is also a college student, probably there for his summer vacation. You do not know how long he has been working but assume it has been some time, since he seems perfectly at ease and occasionally smiles at you. You watch him carefully out of the corner of your eye, trying to get some idea of appropriate "inspecting" behaviors. Occasionally you see him pick up one of the transistors and put it in a bin which you assume to be his reject bin. This increases your uneasiness, for he appears to be earning his money while you have not yet rejected one single part. Meanwhile, the transistors go marching on beneath your magnifying window. Looking around sheepishly, you pick up one of the transistors and throw it in the reject bin. You continue to search for differences in the transistors, but you can not find any. For the remainder of the day, you revert to the practice of indiscriminately throwing an occasional transistor in the bin, trying in the process to appear as confident as possible. Across the way, your colleague looks at you and waves as if to say, "My, how time flies when you have a happy job!" You smile back less broadly and wave to him half-heartedly.

Arriving at work the second day, you are convinced that you will be told of the errors you made. Then you will find out exactly what it is you should be doing. But your supervisor does not appear. The morning is a replay of the previous day. At lunch time, you meet your friend across the way, and chat. During this time you try as casually as possible to introduce the subject of work in hopes that you can glean information concerning the criteria for rejection. He seems indisposed, however, to talk about work, and the time is spent discussing school, football, and, of course, girls.

When you go back to work, you are determined to do your job correctly, so you begin to scrutinize the parts as never before. You notice some of the transistors have one wire longer than the others. Perhaps this is the defect you are supposed to find. You decide that rather than throwing the transistors away haphazardly, a more reasonable approach is to throw away those transistors which have one wire shorter than the other. But soon you find that you are throwing away far too many transistors. At least half of them appear to have this "defect". Thus, you disregard this criterion and again examine the transistors, trying to find the defects which you are being paid to detect. Soon you begin to throw away the transistors on a different basis, and then, discarding that criterion, you select yet another, and still another. Before long, you have exhausted all of the possibilities. There is simply no one way you can detect a consistent difference in the transistors which fits your previous conception of what a defect looks like. Time goes on and you receive neither complaint nor praise for your inspection. By now, you have been working two weeks. It seems ridiculous to approach the plant manager and confess to him that you have been working for two weeks without knowing what you were doing. Certainly, this would not please him. And thus you are committed to a summer of unsystematically tossing transistors into a reject bin, trying to match the rate at which you do this to the performance of your friend across the way and appearing to look interested in your work when, as a matter of fact, you have no idea what your work is.

For most of us, this would not be an enjoyable summer, regardless of the fine wages. We would find that the pretense of taking money for providing a service would weigh too heavily on us. But suppose the job were not just for a summer vacation. It is conceivable that such a situation could go on indefinitely. Imagine how mortified you would feel at the retirement banquet the company gave in your honor after forty years of such deceit; how the gold watch would burn in your hand, the glowing inscription taunt you in mockery! How your stomach would crawl as proud faces of friends and colleagues turn in unison to you and offer strains of "For he's a jolly good fellow!"

How can we possibly help this anxiety-ridden college student whose integrity suffered grievously because of his tender years, his timidity, and his money-grubbing greed? The problem concerned his continued inability to distinguish the defective transistors. It is reasonable that we should ask why the student was unable to detect the difference, and, at the same time. to examine means by which he could be brought to reliably make such a discrimination. Some would suggest that the student was just stupid. We must, however, reject such a conclusion summarily — even the most casual consideration of this proposition would serve to impugn the reputation of our nation's universities.

PRACTICE, PRACTICE, PRACTICE

Another possibility is that the student was inexperienced. After all, he worked as a transistor inspector for a relatively short time. Practice, we have been duly apprised, makes perfect. Perhaps more hours of watching the transistors pass beneath his eyes would have resulted in the ability to discern the faulty transistors.

There is every reason to believe that more practice would only serve to increase the private agony and guilt of the student. If the conditions prevailed, protracted experience would not help in the discrimination of the critical differences. The erroneous belief that practice alone was enough to improve the ability to discern or discriminate differences was first introduced by E. L. Thorndike.[2] You will remember him for his importance in animal research. Later, however, Thorndike reversed his position on the importance of practice, or the "law of exercise", as he called it, by conducting an experiment on psychologists who had come to hear him deliver what was supposed to be his presidential address. Instead of a speech, Thorndike presented the members of the American Psychological Association with pencils and numerous pieces of paper. The audience was told to draw a three-inch line on each piece of paper, and to do so on each subsequent piece of paper. Lines were to be drawn without reference to a ruler, or other means of checking, for accuracy.

The results of Thorndike's impromptu experiment, as well as subsequent data gathered by him and his students, demonstrated that practice, in and of itself, did not help at all. In the case of three-inch lines, the last line was no more likely to be nearer the actual three inches than the first, regardless of how many times an attempt was made.

If practice alone will not help the student to make the correct discrimination, what will? In point of fact, any number of procedures could, even in a matter of minutes, teach a student to detect faulty transistors or a psychologist to draw three-inch lines. For instance, the supervisor could institute a training program for quality-control inspectors. Shock electrodes could be attached to the student's legs or other convenient anatomical areas, and whenever it was apparent that a faulty transistor had passed by the student without being rejected, a shock would be delivered. Doubtless, it would take only a few shocks before the faulty transistors stood out like mountains for the student. In a more positive version of the training program, money could be made to drop from a chute whenever an actual "skag" was rejected. Other less dramatic, but probably equally effective procedures, would be: a movie pointing out the differences between defective and operational components, a speech by the supervisor with or without demonstration, a word or two of explanation from a coworker, notes left by friendly elves, the harping wife of the plant manager verbally abusing each new employee when he makes an error, etc. All these procedures have in common the feature that they affect us, behaviorally speaking, in unique ways.

DIFFERENCES AND CONSEQUENCES

Perhaps we can avoid a long discussion at this point by saying that there is a general statement concerning the process by which one part of the environment becomes differentiated from others.

> DIFFERENCES COME TO EXIST FOR AN INDIVIDUAL WHEN A PARTICULAR SITUATION, OBJECT, OR EVENT AFFECTS HIS BEHAVIORAL SYSTEM IN A UNIQUE MANNER.

A rose is indeed a rose — but only if it behaves as one. If, in fact, it does not smell so sweetly, or if, as the flowers on which Rapachini's daughter thrived, it is fatal to the touch, or if the Rose is the girlfriend of the local Karate blackbelt, then it will be different because it affects us differently.

For our purposes, the environment becomes differentiated because it does not uniformly affect the individual. A kick in the pants is uniquely different from a kiss on the cheek because of the distinctive effects each creates. A doorway is obviously different from a concrete wall, but only because in the past the former has allowed unobstructed passage, while the latter has resulted in a skinned nose or banged knee. An apple is not different from a rock because of its color or shape so much as for the way it affects our mouth and stomach when eaten.

It can be said then, that whenever aspects of the environment provide *advantageous* or *disadvantageous* outcomes, or *consequences*, they emerge as separate and distinct entities. This reference to "advantageous" or "disadvantageous" should not be interpreted to mean that differences come about for the individual only when his tissue is being destroyed through harmful stimuli, or when food, drink, or sex is in the offing. It is apparent that each of us can and does discriminate much subtler differences than these.

A general reference to the basic biochemistry, physiology, and sensory systems of the human, and a discussion of the forms of stimulus energy to which these structures are sensitive, will not enable us to understand the emergence of "awareness". Probably far more important are the individual's previous experiences. Such experiences include special training or instructions which cause given objects to exert unique effects on an individual.

A pile of copper pennies represents to most of us a collection of objects which are essentially alike. Each penny is equivalent to other pennies. The collection is important to us because it can be converted, in stores, to different objects, such as food, drink, or clothing, which have immediate, differential significance to hungry, thirsty, or naked men. To the coin collector, however, each penny may create such distinct effects that it is a universe in itself.

The fact that each of us experiences things in a unique way is the result of an interaction between our unique biological systems and our unique experiences. Entities continue to exist for us because they continue to bring specific outcomes to us, outcomes which either harm us or promote our survival. The agency which delivers these outcomes can be as impersonal as a bolt of lightening, or it may involve the highly personal intervention of a parent who rewards us when we behave appropriately toward differences he wishes us to discriminate.

Early in our lives, the most significant unique outcomes are brought to us by human agents. By coordinating their words or deeds with the differences they wish us to discriminate, those differences soon begin to emerge for us and to affect our behavior in unique ways.

In many cases differences can be taught to us with little effort on the part of teachers, supervisors, or parents. In other instances, such as art, drama, or skills involving adroit, coordinated movements, differences become apparent only after years of persistent study in the presence of a keenly sensitive tutor or coach. Initially, it would appear that we all need something or someone to tell us when we are right or wrong. When this happens consistently the intermediary may cease to be necessary, and we can discriminate differences and judge the appropriateness of our behavior independently of supplementary cues.

It is perhaps possible to see now that in a very real sense other people create differences for us. By applying their words or actions consistently, they bring us to make discriminations and discern differences which otherwise would never have existed for us.

DIFFERENCES AND REALITY

But, if other people by their words and deeds actually create differences for us, does this necessarily mean that these differences are *real*? Remember, the incestuous episodes that Freud helped create for his patients, and the beasts that Thurber sketched were ultimately judged *unreal.*

Only a brief glance at history or a cursory examination of contemporary superstitions should convince you that demons, ghosts, unidentified flying objects, animal spirits, the Abominable Snow Man, and Santa Claus do exist in the world of some individuals. These entities seem to be as palpable to some people as a brick wall, or the impact of a bruised shin, and for the most part, they came into existence because other individuals provided consistent consequences for them. The systematic application of consequences, in the form of punishment for behavior inconsistent with a belief and reward for behavior consistent with it, defines what we commonly call teaching. Again, it must be stressed that *the procedures by which entities or differences come to exist for an individual are separate from the procedures by which they are judged real or unreal.*

DIFFERENCES AND MEASUREMENT

We are still pursuing the question, "What is measurement?" A common reply takes the form of: "Measurement is a procedure for discerning differences among objects, things, or events." The existence of the objects, things, or events is assumed, or taken for granted. Measurement is merely a procedure for finding out if objects are different or not. When two objects are related through measurement operations to a third object, a ruler, for instance, differences become apparent. But wait! Measurement is not always a necessary step in finding differences. Surely rulers, micrometers, or calipers are not required to discern that a boy is different from a girl or a rock different from an apple. In these instances, differences are readily apparent. By what procedure are these differences ascertained, and how does it relate to procedures utilizing measuring instruments?

A brief look into our childhood reveals a time when we did not in fact distinguish a boy from a girl or a rock from an apple. Our parents made these differences exist for us. By applying differential consequences to our behavior with respect to its appropriateness in the presence of these objects, differences became apparent to us. Learning these differences early was inexpensive relative to outcomes which society and our physiology would, undoubtedly, provide later if they had not been learned.

Thus differences exist for us prior to our experience with rulers or other measuring devices, because our general environment has behaved in very consistent ways toward our interaction with it. A brick wall teaches us that we cannot pass through it because we bump our nose or bang our shin when we try to do so. It does not need the help of parents to provide outcomes. It has its own consequences which become immediately relevant to humans who do not discriminate it from the open doorway.

People who conform to the rules of society may respond almost as consistently as the brick wall toward those who behave in ways which go against the "natural" laws of society, inasmuch as they consistently provide consequences for children in many relevant situations. From this point of view it may be seen that the devices and procedures which are involved in what we call "measurement" are in reality nothing more than systems for creating differences. In some cases measurement creates easily discernible differences for large groups of individuals. Initially, these differences may have existed for only a few. Measurement makes them exist in the common experience of the culture.

For most of us, minor differences in the length of boards do not exist, but these differences have come to exist for the carpenter because of a long "training" period in which he has been punished for cutting boards too short, or too long, and rewarded for accurate cuts. Fortunately, we do not have to go through the same long experience of the carpenter before these differences will exist for us. They come into existence immediately when we apply a ruler. This procedure is, of course, called measurement. It creates differences where previously none were apparent.

The highly skilled musician may be competent to discriminate a tone which differs from another by only a small frequency without the use of visible measuring devices. But by applying the correct measuring equipment this difference becomes visible as the deflection of a meter or the excursion of a graduated dial which can exist for all. Thus, we can discriminate as well as the musician, and without years of practice or training.

In some cases, measurement makes differences exist which previously did not exist for anyone. In Chapter 1, the discovery of the positron, or positive electron, was clearly the creation of an entity which previously was outside human experience. The measurement procedure employed by the technician created the positron for Dr. Anderson. He in turn created it for the scientific community.

MEASUREMENT AS A CREATOR OF DIFFERENCES

Let us proffer a definition of measurement which can be used consistently throughout this book and will promote your understanding of psychological tests and measurements. To summarize the discussion to this point:

1. Differences resulting from comparison of successive experiences in the same situation, or simultaneous comparisons among a collection of objects, come about for a particular individual because unique consequences are consistently applied to his behavior.

2. These unique consequences are applied by other individuals in the culture, and thus are part of the socialization, training, and educational processes of the culture; or they may be a "natural" result of the individual's interaction with the environment itself, independent of the culture.

3. Unless the behavior of the individual comes under the influence of these unique consequences, these differences will not emerge for him. They will not exist.

4. Procedures usually referred to as "measurement" translate differences which exist for only a few into a medium which will make these differences exist for others.

5. In some instances, a new set of "measuring" devices or procedures causes new entities, or differences, to exist where previously none existed for any individual.

MEASUREMENT: A DEFINITION

It may have occurred to the reader that the conceptualization of measurement developed in this chapter conflicts with common sense notions. Therefore, let's spend a little more time trying to clarify and justify our position. Consider the following definition: MEASUREMENT IS A SET OF PROCEDURES AND DEVICES FOR THE CREATION OF DIFFERENCES. Viewing measurement this way may seem foreign to our intuitive understanding of the term. Clearly, when we think of measurement, we all too quickly insist on an object, thing, or phenomenon to be measured. This tendency arises from our experience with rulers and other devices which are directly applied to objects which have already become differentiated from the environment. We know a board exists *prior to* placing a ruler on it. It exists because we can see it, knock against it, and feel it. Its effect on us is direct. Differences defined by a ruler do not seem to come into existence only as a consequence of measurement procedures.

"Measurement does not 'create' differences, the reader may protest, it merely detects differences which are already there!"

But, as we emphasized in Chapter 1, the procedures whereby differences come to exist are *separate* from the procedures whereby we decide whether they are real or not. It is worthwhile to point out also, the elementary truth that not all differences are important. More correctly, as we shall learn later, it is safe to say that only some differences are important for some purposes.

Consider for a moment a device which a crackpot inventor has just constructed. It is laid on the human body in a particular manner and subsequently deflections appear on a meter at one end of the device. Is this device enacting measurement? If so, what is being measured? It is impossible to answer these questions. A more important question would be: Are the differences observed, which are created by the device, of any value? If, in fact, the differences ultimately prove to relate to other events or differences of known value which have been created by other devices, then the device is valuable. Furthermore, none would question that it could rightfully be called measurement.

In future chapters, the advantage of this definition of measurement will become apparent. Intelligence, personality, or potential, unlike a board. cannot be seen, knocked against, or felt. Differences among individuals in these areas are more obviously created by measurement procedures. In the next chapter, psychological tests will come into focus, and their peculiarities will receive attention.

FOOTNOTE

[1] Thorndike, E.L. *The Fundamentals of Learning.* New York: Teachers College, Columbia University, 1932.

CHAPTER 4

Psychological Tests: A Definition

In the last chapter a definition of measurement was presented which should have caused the reader at least mild concern. The definition was: *Measurement is a set of procedures and devices for the creation of differences.* Since a psychological test is a form of measurement, the procedure by which it is administered constitutes *a means for the creation of psychological differences.*

"But the procedure and the test can't possibly be all there is to testing," you may say, and you are right. So let us examine a little closer some of the means by which psychological differences are created and also focus on events which could influence or alter resulting measurements or test scores.

THE NAKED PSYCHE

A nervous young man paces through the waiting room of a psychologist's office. He is visibly disturbed, and except for the sake of his health, he might be smoking one cigarette after another. Now and then he sits down, thumbing through magazines, but not really exhibiting any interest in them because he is preoccupied with the office door and irritated by the receptionist's placid indifference. Eventually, the door opens and a conservatively dressed man in his late thirties calls the young man by name. "Please, come in, Mr. Doe," he says.

The young man follows him in and sits down in a large overstuffed leather chair. The office is comfortably appointed with rows of books and sturdy, modestly unobtrusive furniture. The psychologist looks out of the window as late-afternoon sunlight streams through the vented blinds. After a moment, he turns and takes a seat in the swivel chair behind the large desk. He reaches into the bottom drawer and removes a packet of cards which, with a notepad and a pamphlet containing figures or designs, he carefully arranges before him. Leaning slightly forward, he expertly explains the task. "I'm going to show you some pictures which are on these cards. They were made by dropping ink on a piece of pape and folding it. I will show you the pictures one at a time, and I want you to tell me what you see in the designs." The young man's hands tighten nervously, and there is a dryness in his throat as he takes a deep breath. He is relieved that the desk conceals the slight trembling in his legs.

"What do you see here?" A card is held toward Mr. Doe, who begins to reach for it automatically, but then abruptly retracts his hand and places it self-consciously in his lap. Feeling somewhat foolish and expecting the worst, he forces himself to look at the picture. A black design sprawls nondescriptly over most of the card. "Good grief", he mutters, relieved at the apparent innocence of the design; and overwhelmed with acute embarrassment he tries to decide on a reasonable course of action.

. . . .

The reader should not interpret the foregoing account as entirely hypothetical. Psychological testing continues to evoke anxiety in a major segment of the public. The psychologist and his craft remain mystical, his tests incomprehensible and mysterious. Undoubtedly much of the mystique of psychological testing is needlessly engendered by those professionals who construct, administer, interpret and publish psychological tests and test materials. One highly respected professional has called the Rorschach (inkblot) test "an x-ray of the psyche".[1] Understandably, few of us would take lightly the laying bare of our psyche.

Much misinformation and sheer myth is actively perpetuated by professional groups who, unwittingly, work to keep the lay public in ignorance and awe of the psychological test. Just as the secrets of sex must be secure from the young, so must the secrets of psychological tests be kept from those who would irresponsibly harm themselves or others with them. Or so they claim. Some professionals even urge stiff fines and jail sentences for unauthorized possession of test materials. Imagine intercepting the following police call:

> "Calling all cars; calling all cars. Man at corner of First and Elm with 'loaded' intelligence test kit. Proceed with caution. Suspect allegedly armed with concealed Rorschach cards."

Is such an extreme position warranted? Indeed, there is an objective need to restrict the general availability of test materials if their usefulness is to be preserved. Who would want a society that diagnosed its own neuroses and tested its own intelligence? Obviously, we would wind up in a world where everyone old enough to read would be super-sane and super-smart, thus eliminating the tremendous gratification man usually gets from knowing that he is superior to his wife or best friend. He would have to look down at infants, illiterates, the blind, and foreigners to retain his "holier than thou" attitude. And of course we would have to retain the droves of professionals who used to engage in the awesome service of pinning labels onto our psyches.

Viewed in this light, the reader will have to agree that certainly harm could be done to mankind through the improper use of tests and test results. But inestimably more harm could be done by the deliberate secrecy perpetuated by certain professional groups who would [illegible] to see an unquestioning public [illegible] through awe, ignorance, and intimidation.

We cannot promise to rid you of the awe and intimidation; but we will certainly try to remove your ignorance as to what a psychological test is or is not, and what it actually measures. But before illuminating this whole affair for you, let us consider some of the difficulties of measurement.

THE BLOCK OF STEEL

Propose that a thousand different individuals measure the length of a block of the hardest steel. Provided the ruler has finely graduated markings on it, perhaps in ten-thousandths of an inch, it would be unlikely that any two of the thousand people would arrive at the same exact measurement of the steel. Prior to Albert Einstein,[2] it was assumed that objects existed independently of their measures. A block of steel had a definite width, height, and length. These exact or "true" dimensions were never directly accessible to mere mortals. Presumably, Mother Nature kept them locked up in a giant ethereal fortune cookie somewhere.

Try to reconcile the assumption of "true" dimensions with our thousand measurers and their thousand different measurements. Which of these measurements is correct? It is certain from this point of view that since all measurements are different, only one can be correct. But which one? Will the real measurement please stand up! What a pity we cannot open the fortune cookie and end the suspense. Of course, there is another possibility, namely, that none of the measurements are correct. But there is no way to tell. Once again we must lament the inaccessibility of the fortune cookie.

Einstein's unique contribution was to introduce a third possibility. All of the thousand measurements are correct! Einstein's contention was that Mother Nature's fortune cookie for the moment was private property; it did not appear that she was willing to share it with us. The problem boiled down to a practical consideration. It was no longer reasonable to keep asking, over and over again "which measurement is correct?" "Correctness" was not the issue. We are not concerned with correctness in terms of the correspondence of our measurements to some absolute and perhaps imaginary standard somewhere. We are interested in measurements which work. Correctness was not the criterion by which one measurement would be more highly valued than the rest. We would use the one which worked best for us. "Best" would further be qualified in terms of specific outcomes or criteria.

THE CARPENTER'S APPRENTICE

Measurements created by a ruler continue in use because they work for us. With the ruler even a weekend carpenter can be relatively secure in cutting boards which will fit properly. Of course, it is possible to use a ruler in such a manner that measurements or differences created by it will not work. It must be held in a standard way before it can be of consistent value in building a bird house or panelling a spare room. Thus, if a ruler is employed in a standard manner by different people and at different times, we will find that while agreement among them may not be absolute, the slight differences are such that a carpenter will not feel too apprehensive about having the staff at the lumber yard cut the pieces to his specifications. The carpenter has found that the staff is trained sufficiently enough in the use of rulers that the boards they cut fit as well as those he himself might measure and cut.

But suppose the carpenter takes on an apprentice who, because of innate inability, moral laxity, or sheer perverse nature, repeatedly holds the ruler in different ways and introduces new and creative interpretations of the markings on the ruler. We would not expect the carpenter to even consider using the boards cut in this fashion, for they would not function for him. Much expensive lumber and valuable time would be wasted. The carpenter would be quick to deplore the apprentice's worthless measurements. In point of fact, this judgment reflects the very specific purposes for which the carpenter traditionally employs measurements. It is certain that measurements made by the apprentice are related to other factors, but these are of no interest to the carpenter.

Measurements made by the apprentice will, however, tell us a great deal about the apprentice himself. Earlier in this chapter we presented a hypothetical encounter between a young man and his first psychological test. As most of you already know, the Rorschach test which the young man took consists of designs produced in such a way that they do not obviously resemble any particular object or set of objects. The inkblots are deliberately chosen to present the subject with *ambiguous* visual stimuli. When he is asked to relate what he sees, his verbal report, and other general behavior, tells more about him than it does about the inkblots, and this is precisely the psychologist's aim.

The carpenter, however, is upset with the measurements produced by the unique interaction between the ruler and the hapless apprentice. The apprentice's purpose is to cut boards which will fit, so the carpenter watches the apprentice measure the board while the psychologist watches the client measure himself.

Thus, the subject interacting with a psychological test is not appreciably different from an apprentice interacting with a ruler. Both involve a human being interacting with devices, procedures, and other aspects of the environment. But more importantly, both situations create differences, or measurements. In one instance, these differences help the carpenter predict how boards will act when they become involved with notches and joints and grooves in their environment; in the other instance, differences or measurements can help the psychologist predict how an individual will act in school, in work, in his interpersonal relations, and so forth.

FEATURES OF PSYCHOLOGICAL TESTS

It is not always simple to clearly distinguish psychological differences or measurements from other types of measurements, so let us concentrate for a while on those "difference-making machines" termed "psychological" and try to extract some common features which distinguish them from other types of tests.

1. Psychological tests do not involve devices which go beyond the skin. Such tests as those employed by the physician or physiologist may entail entrance into the body of the subject. In some cases, a psychologist may also be involved in research which utilizes such measures. However, psychological tests, as they are currently available to practicing psychologists and educators, do not invade the subjects' bodies. In fact, in most states it is illegal for a psychologist to violate the skin while offering a service to the public. (What goes on after office hours is an entirely different matter.)

2. Psychological tests employ the subject as an active agent to produce differences. They do not involve differences in the hardness of the subject's bones, the amount of pigment in his eyes, or the number of fingers or toes he possesses. Psychological tests, in interaction with the behavior of an individual, create differences which were not previously observed. It is the subject's behavior — what he does — to which a psychological test is applied. Thus, it creates measurements of the behavior of many subjects, or of the same subject on different occasions. To the casual observer, all of the subjects are emitting the same behavior. They are all diligently looking at the questions on the test form and making marks in the test booklet. The psychological test, however, will later allow us to describe differences in the performances of these individuals. Just as two boards may forever remain "the same length" until they are measured, so may individuals seemingly behave in the same way until a psychological measurement is made.

3. In psychological testing, the differences which are created are generally expressed as numbers or scores. Indeed, most measurement, as we know it, involves differences which are expressed as numbers. Perhaps psychological tests differ from other forms of measurement in the sense that sometimes the differences appear in other than numerical language.

Measurement need not necessarily involve numbers. As a difference-making procedure, measurement may create differences by means of any language system, including pig latin. But a procedure which relates differences in numerical form is superior, because more differences can be created by it than by one which does not use numerical language. Furthermore, numbers can be related to one another and manipulated in various ways. For these reasons, numerical language is far superior to other language forms, such as words or symbols.

The most distinguishing feature of the psychological test is its purpose. The purpose of all psychological tests is to *predict* how the individual will behave in the future. Psychological tests allow for prognostication. They enable us to leap ahead into the future and make a qualified guess as to what a particular person may or may not do at that time. In a sense, the psychological test allows us to become time-travelers. We may pass into the fourth dimension without the necessity of all the gadgets and paraphernalia described by H. G. Wells in his book entitled *The Time Machine.*[3]

There is no real way to tell which particular situations, observations, or measured differences will serve as predictive indices for future behavior or performance. Later, when the topic of validity is considered, this will be discussed in greater detail. In order to give you a hint, however, let us at this point discuss one currently relevant situation.

In the past, graduate schools have been hard-pressed to find psychological tests which would help them to select those students who were most likely to utilize the limited available openings. One prevalently used test, the Graduate Record Examination (GRE), has been found to relate poorly to success in graduate school, and a recent study has shown that the GRE relates to success as a psychologist even less.[4]

In the meantime, one researcher[5] has hit upon what appears to be a very promising test. It is, of all things, an *eye* test! This researcher found that individuals suffering from near-sightedness, or myopia, do significantly better in graduate school than other candidates do. Furthermore, the eye test is an infinitely better predictor than the GRE, or even an index compiled from undergraduate grades and a series of traditional tests. But why should near-sightedness relate to success in graduate school? At this point, the answer will be available only through more research.

Now, to fulfill the promise of this chapter, let us define a psychological test in the following way:

A SET OF STANDARD STIMULI AND PROCEDURES, INCLUDING ENVIRONMENT, INSTRUCTIONS, PRESENTATION OF TASKS, AND SCORING CRITERIA, WHICH, WHEN APPLIED TO AN INDIVIDUAL, YIELD STATEMENTS WHICH CAN BE USEFULLY RELATED TO HIS FUTURE BEHAVIOR.

FOOTNOTES

[1] Graham, Ellen. "Now the Boy Inkblots Look Like Girl Inkblots and That Savs a Lot." *Wall Street Journal.* August 20, 1971. p. 1. Cited by Dr. Fred Brown at New York's Mt. Sinai Hospital.

[2] Slossen, Edwin E. *Easy Lessons in Einstein.* Harcourt-Brace-Jovanovich, Inc. New York, 1920. p. 97.

[3] Wells, H. G. "The Time Machine." In *Seven Science Fiction Novels.* Dover Publication: New York, 1950.

[4] Marston, Albert R. "It Is Time To Reconsider The Graduate Record Examination." *American Psychologist.* Vol. 26, July, 1971. pp. 653-655.

[5] Young, Francis A. University of Washington, Pullman, Washington. Cited from a personal conversation in October, 1971.

CHAPTER 5

Test Results and Numbers

In the last chapter, a definition of a psychological test was finally introduced. It is likely that you have already arranged to have it cast in bronze and mounted in your den, bathroom, or wherever you devote yourself to hard thought. But just to be on the safe side, let us pause to refresh you. A psychological test is:

> A SET OF STANDARD STIMULI AND PROCEDURES, INCLUDING ENVIRONMENT, INSTRUCTIONS, PRESENTATION OF TASKS, AND SCORING CRITERIA, WHICH, WHEN APPLIED TO AN INDIVIDUAL, YIELD STATEMENTS WHICH CAN BE USEFULLY RELATED TO HIS FUTURE BEHAVIOR.

For the present, we shall be concerned with the last clause of the definition, which refers to "... statements which can be usefully related to (an individual's) future behavior."

When a psychological test is given, test materials are introduced to an individual in a standard manner along with standard instruction. In responding to this overall testing situation, the subject's performance creates test results. These results are the "statements" spoken of in the definition above. To the extent that these statements or results can be "usefully related to his future behavior", the psychological test has merit.

It is time to consider some characteristics of a psychological test which tend either to limit or enhance its *predictive power* — the extent to which it can be related to future behavior. By way of illustration, imagine someone less noble and scrupulous than yourself, someone whose base motive is to make a quick financial gain in the psychological testing business.

WHAT MAKES SAMMY SLICK?

Onlooker: Gosh, Sammy, fantastic! You are making an intelligence test right before my very eyes. I didn't know you had a Ph. D. in psychology and belonged to the American Psychological Association and all that. Golly! And you're barely 47 years old!

Sammy: Don't be a dunce, you dunce. I got most of the test items from Reader's Digest. Besides, I subscribe to Psychology Today. There, it's finished. Now, on to make a million!

Onlooker: How exciting! A real intelligence test! Just like Binet and Stanford, Wechsler and Bellvue, and all those other guys!

Sammy: Can the chatter. Let's get on with the money making. Is that kid still waiting?

Onlooker: Oh, yes. His mother has him outside. She is so eager to find out how smart he is. But, are you sure your test will work?

Sammy: Don't worry about that now. Quick, sit the kid down. Here, look at this, kid.

(2½ minutes later.)

Sammy: Well, that's that.

Onlooker: But Sammy, only 2½ minutes? Other tests take an hour or two to administer.

Sammy: Yeah, they don't make money very fast, either. This is the short form of the "Sammy Slick Smarts-Selector Survey". Let's go see the kid's mother. (Later, with mother.) Well, Mrs. Stanley, your youngster Clifford, here, has an I.Q. of 437.

Mrs. Stanley: Goodness, isn't that pretty high?

Sammy: I think you could say your son is a genius ... yes, a genius.

Mrs. Stanley: My boy! My boy!

(2,000 kids later Sammy and the Onlooker are in Sammy's new plush office. The Onlooker has some tests in his hand.)

Onlooker: I don't understand, Sammy. We've tested thousands of youngsters since you first constructed your test, and all of them have achieved I.Q.s of 437. Isn't it odd that all of them should be so intelligent?

Sammy: Remarkable, but not surprising. The "Sammy Slick Smarts-Selector Survey" is one helluva test!

Onlooker: But wait, Sammy. I was just looking at this test. If you even lift your index finger, the only possible score you could get on the test is 437. A two-toed sloth, for instance, tests out the same. This means everyone who takes the test will get exactly the same score. There's something wrong here, Sammy. A test like that can't be any good!

Sammy: Oh, yeah? Have you followed me to the bank lately? Besides, my test measures "true" intelligence; it's completely free from measurement error. If the government comes through with the money they've offered me, we'll completely wipe out low intelligence by eliminating measurement error.

Onlooker: I'm sorry, Sammy, I should have realized: but what will that mean?

Sammy: The end of poverty, crime, injustice, bigotry, violence, pellagra, dishonesty, itchy scalp, and venereal disease.

Onlooker: Sammy, I'm ... well ... honored to know you. And say, Sammy, would you, could you ... do you suppose I could take the test?

FLEXIBILITY OF THE LANGUAGE SYSTEM

In the preceding account, the "Sammy Slick Smarts-Selector Survey" was of questionable worth to anyone but Sammy. The statements which resulted from administering the test could not possibly be related in any sense to the future performance of the subjects. There was only the single outcome of 437, which was attained by all takers of the test. Since I.Q.s produced by Sammy's test were all the same, there was no possible way these statements could be related to future behavior. Obviously, before any possibility for such relationships can occur, the test must yield different results for the individuals who take the test. A judge who gives all Miss America contestants the same score does not help in selecting the most beautiful, although he may find his work enjoyable. A sex-appeal test which reflects the same score for Phyllis Diller and Raquel Welch would be a risky device to employ in selecting a date for the big weekend. But probably the worst indictment against Sammy is that his test did not fit the venerated and sanctified definition of measurement which we introduced in an earlier chapter. If measurement is a set of devices and procedures for the creation of differences, then Sammy's test does not qualify for membership in the Measurement Club because no differences were ever obtained.

As a first requirement, separate statements derived from the application of a psychological test must be different to some degree before they have any possibility of being related to future events and before they can truly fit the definition of a psychological test. The *Sammy Slick Smarts-Selector Survey* is not an instance of measurement, nor is it, in terms of our definition, a psychological test. (Sammy can be heard sobbing on his way to the bank.)

But before we dispose of Sammy Slick's efforts, suppose that instead of a flat I.Q. of 437, Sammy had been a bit slicker and designed a test which came up with *two* outcome possibilities, such as "smart" and "stupid". Sammy could now be accepted by our elite corps. Differences are created by the test, since some individuals are found to be "smart" and others "stupid". Sammy's test qualifies, by our definition, as a psychological test and its application as a bona fide instance of psychological measurement.

It is possible that results on Sammy's test could now be related to the future performance of his subjects. Those who achieve a result of "smart" may do better in school, business, and life in general than those who receive a "stupid" rating. But wait! Couldn't finer relationships be possible if a third category, such as "normal", were added? Now with the three categories — stupid, normal, and smart — wouldn't Sammy's test be better? Actually, we cannot say if the test would be better until we have additional information. We can definitely say, however, all other things being equal, that there is a *greater possibility* for relationships between test results and future performance to come about when there are three categories instead of two. According to this line of reasoning, other categories, such as "really stupid" or "really, really stupid" or "abysmal", can be added. Through the use of the common English language, replete with superlative and diminutive modifiers, an immense number of categories could be generated or created by the test, thus increasing the possibility for relationships with the future to be established. Whether or not such relationships would occur is entirely another matter.

For the moment, let it be said that the greater the number of different statements which can be generated by a psychological test, the greater the possibility, all else being equal, of establishing relationships with future events. Thus, by this criterion, such a test is *better*.

But rather than using cumbersome modifiers taken from ordinary language to increase statement categories, a more effective approach is to adopt measurement procedures which yield statements in *numerical* language. A measurement system which yields statements such as 25, 49, 347, etc., is much more convenient than one which generates statements such as "really nice", "hunky-dory", "peachy keen", or "marginally repugnant". But do not be mistaken in the belief that the utilization of a number language in and of itself enhances the possibility for the discovery of relationships between the test result and future performance. The substitution of two numbers such as "1" and "2" for statements of "pass" and "fail" or "smart" and "stupid", is of no advantage whatsoever. Any limitation inherent in the language system in which statements of test results are reported will reduce the likelihood that such outcomes can be related to future situations. Tests which yield pass-fail results are more likely to be inferior to tests which yield percentage scores. By the same token, however, tests whose results are expressed in percentages have less potential than tests whose scores are not limited to a ceiling of 100%. A professor who gives examinations where all students score 100% will find it impossible to employ those exam scores to select the student most likely to succeed in graduate school.

If you have been following closely, it has probably occurred to you that the test with the greatest possibility of success in predicting future performance would be one in which test results or statements occur in a language system which is entirely unlimited. Indeed, you are correct. An excellent one would employ the natural number system we all know and love, which as you well know goes 1, 2, 3, etc., and would extend to infinity. Naturally, infinity is a theoretical limit. In reality, all tests have an upper limit. It is still true, however, that the less restrictive the upper limit, the greater the possibilities of the test.

Up to this point we have been talking about some characteristics of language and have indicated that it is impossible for some kinds of differences to occur in certain languages because of the inflexibility of those languages. A test whose results are expressed in a language where a wide range of distinctly different outcome statements or test results are possible possesses a greater potential for being related to other events, measurements, or observations. It is one thing, however, for outcomes to be expressed in a language where differences *could* occur; it is quite another thing for a wide range of differences to *actually* occur when the test is administered to a group of individuals. Hopefully, the following piece of science fiction will elucidate this point.

EXCURSION TO THE PLANET BI-NO

For decades, sociologists and anthropologists have journeyed to remote parts of the world to study the behavior of various cultural groups in their natural habitats. As space travel becomes a reality, future social scientists will perhaps travel to distant solar systems and, in a similar fashion, study the behavior of extraterrestrial creatures. Imagine that in this future time, one such individual — a Buck Rogers' version of Margaret Meade, voyages to the planet BI-NO for this purpose. He takes with him many psychometric and anthropometric devices, among them a gauge designed to measure strength of grip. On Earth, the grip gauge was an excellent instrument. Differences in grip ranging from a few grams to several tons of pressure could be read from the indicator.

Upon arriving at the planet BI-NO, our space scientist discovers some of the peculiarities of the planet. There is no dawn or dusk — just day and night; no breeze, but hurricane or dead calm; no gradations of temperature — only blistering heat or freezing cold. In fact, things all seem to be one way or the other with nothing in between. The inhabitants of BI-NO also reflect this dichotomy. They are either extremely short and plump or tall and gaunt; either smooth as billiard balls or covered with shaggy hair. But more to the point, they are either so weak that they quiver at the sight of their own shadow or so powerful that they must exercise great care lest they smash anything they touch. When the BI-NOs are given the grip test, two scores are seen: either zero, indicating a grip too slight for the gauge to register, or the maximum, indicating a grip of more than ten tons. The grip test, which was very useful in detecting a great range of differences among earthlings, is comparatively useless with the either-or population of BI-NO.

In a later chapter we will discuss the problems encountered when a psychological test developed for one group of individuals is later employed with another, quite different population. For the moment, a summary may help clear up any residual confusion.

1. A test which generates results or outcome statements in a language which limits the number of discretely different statements which can possibly occur restricts the likelihood that the test will be of value in the prediction of future outcomes.

2. A test which generates results or outcome statements in a flexible language, will still be restricted if few differences actually occur when the test is applied.

Both statements above are reducible to the generalization that, all other things being equal, the value of a test is limited by the absolute number of discrete differences generated when it is applied. As a difference-creating machine, the best test will probably be the one which actually creates the largest array of observed differences. It is this test which has the greatest chance of helping us to predict an individual's future behavior.

It should now be less of a mystery to you why measurement should generally result in numbers rather than words such as "regular", "long", or "king-sized". A test which creates differences in numerical language is convenient, familiar, and, more importantly, acts in no way to limit the number of discrete outcome statements possible.

There is, of course, another reason for the use of numerical language in preference to common language. It is the fact that numbers can be manipulated and transformed in a variety of ways. This is accomplished by means of a most ingenious and mystifying set of procedures called mathematics.

WHERE DID NUMBERS COME FROM, DADDY?

Perhaps, like the author, you are one of those who have great respect for numbers but are absolutely panic-stricken at the prospect of having to perform or interpret mathematical operations. There is, after all, an arcane mystique associated with the practice of mathematics, a certain inscrutable sorcery implicit in the craft of the "mathemagician". It is not surprising that the high priests of the number are often smug, insensitive, pedantic and snobbish. What is worse, and entirely unforgivable, is that they appear to be infinitely more intelligent than we are. While they glide effortlessly through life on a well-oiled slide rule, the rest of us must finger and toe it as best we can. But aside from the wizardry and injustice permeating mathematics, one question remains an eternal mystery where numbers are concerned — where did the damned things come from in the first place?

There is some degree of face-saving in the knowledge that even the early Greeks (who, you must admit, really were a bunch of know-it-alls) also had trouble with this one. We might add that their attempts to answer the heavy question of the origin of mathematics were undoubtedly responsible for the magical aura of the subject, an impression that haunts us to this day. For instance, regardless of how keen we are on Pythagoras' old hypotenuse, it seems he went a bit far when he pondered that "number is the pervading reality of all life and substance".[1] Nonetheless, generations of number freaks have followed to perpetuate the mysterious, even religious quality of numbers. Today the magic is most apparent in the machinations of numerologists who add up the letters in your name or count the number of hairs on your head and then make such awesome predictions as, "You shall meet a dark stranger who will change your life", or "You shall grow older before long".

Regardless of what mathematicians and numerologists, either classical or contemporary, would imply, numbers and mathematics are man's invention. If they are to be understood, their advantages to man must be given foremost consideration.

As perhaps the simplest operation with numbers, "counting" is a good place to start in the quest for understanding. Whether it involves the Roman numeral system, the Arabic system we customarily employ, marks in the sand, or the old fingers and toes, counting behavior is not significantly different from other *verbal behavior*. It originally developed and still remains active in our individual and cultural response patterns because, like all verbal behavior, it creates definite, advantageous outcomes for us.[2]

Thus, if one would inquire as to why man began to count, he will find the answer by asking why man began to talk or write. A number is no more or less than a word. Certainly if individuals were for any historical or biological reasons likely to band together in communities, cooperation at least of a minimal sort was essential. Initially, talk made living together easier — a conclusion which is perhaps not obvious to many contemporary husbands and wives.

It is unlikely that man's first word was a number.[3] However, if it was mutually advantageous to alert one's tribesmen to the location of game or the enemy, it was soon obviously advantageous to include a *quantitative* description. Just as early man would behave differently when preparing for an encounter with the animal called "bear" than he would for the animal called "sheep", so would the word "six" control different behavior from "twelve", "twenty-one", or "one-hundred".

Anthropologists are not in agreement as to whether non-numerical language began in a spoken or written medium,[4] but there does seem to be a consensus that numerical language began as written communication, probably in the form of marks on the earth or scratches on stone surfaces. Marks "stood for" things. Later, more convenient and flexible notation systems evolved, and written numbers were given spoken counterparts.

Counting is one thing and mathematics entirely another. The former by no means necessitates the latter. A child may say "his numbers" as a rote exercise. For him numbers have no further meaning than the vocalization of sounds which are followed by pats on the head, hugs, and the general appreciation of his elders. In a similar fashion, youngsters learn to write a consecutive chain of numbers, or respond by writing a given number on command. None of these are "counting", although they may be called that. Counting occurs when numbers correspond to or "stand for" objects. Asking a child to count beans in a jar is different from asking him to recite the numbers in the absence of specific stimulus objects.

Two separate piles of stones may be counted by making a mark in the sand for each stone. This procedure will produce two separate clusters of marks, each cluster corresponding to one of the separate piles of stones. If for some reason there arises an advantage to knowing the size of a grand pile of stones composed of the two smaller piles, one pile may be carried stone by stone to the other. It is difficult to say how long it took man in this "stoned" condition to discover that the same result could be achieved by transferring one cluster of marks to the other. With this discovery came "addition" and the beginning of mathematics.

Operations involved in simple mathematics are really nothing more than physical manipulations performed on numbers and marks in much the same way as they had been previously performed on the objects with which the numbers were associated. Terms which currently occur in mathematical instruction, such as "take away" and "carrying", are vestiges of the "natural" relationship between the manipulation of numbers and manipulation of objects.

The advantage of manipulating marks on sand or slate as opposed to the manipulation of real objects is not apparent where the traditional oranges and apples are concerned. However, when the objects are wild bulls, love-starved gorillas, or radioactive isotopes, or where the number involved is in the thousands or tens of thousands, it is not surprising that the marks win out. But what probably began as convenience gradually evolved into a miracle few fully appreciate. As the advantage to manipulating objects too huge, too remote, or too dangerous became apparent, mathematical operations surpassed man's ability to perform the physical operations on which they were based. Although theoretically possible, such physical operations were in fact beyond human capability. What all of the king's men and horses could accomplish only in a million years, a lone, spindly-legged scribe, slate in hand, performed in a twinkle.

Even so, simple mathematical operations do not cause general confusion. The postulates which go into the natural number system (i.e., addition and multiplication, as a special case of addition; subtraction and division, as a special case of subtraction; and radicals and geometric progressions as special cases of division and multiplication) are easily taught as real-life concepts. Indeed, this is the premise on which modern math is based. A severe problem in understanding arises, however, when negative numbers, non-Euclidean geometry, or imaginary number systems in general are pondered. One who was previously impressed by the uncanny correspondence between mathematical operations and physical operations becomes quickly dismayed at his inability to find a place in the world where the physical operations corresponding to $\sqrt{-1}$ occur, or where parallel lines converge. When he is told to accept as obvious that the $\sqrt{-1}$ is critical in predicting the behavior of the electric current which runs his television set and that the redoubtable Einstein demonstrated the meeting of parallel lines as a fact of the universe, dismay becomes a feeling of abject stupidity and loss of self-worth.

Again the question returns: From where did these formulations spring? Since many mathematical operations have no physical-world counterparts, they cannot, as was the case with simple mathematics, exist as symbolic representations of physical operations.

There is a clue in the observation that they are called "imaginary" number systems. Even as you read this sentence, mathematicians huddle in smug-filled rooms dreaming up recondite formulations which will baffle future generations of college students. These theoretical mathematicians are not, strictly speaking, concerned with the real world. They are not in the least deterred by the fact that the assumptions or postulates by which their formulations are generated could only be experienced in Wonderland, Oz, or Hell. If they will forgive a somewhat unflattering analogy, it might help if you consider the theoretical mathematician as a mad tailor who from day-to-day designs a succession of new and fanciful garments. These garments are not designed for any existing creature; some have holes for eight or nine heads, or 27 legs, but no arm openings or sleeves. Others appear designed for creatures with small bodies and gigantic heads, or for organisms with huge bodies and tiny, pin-shaped legs. As each day passes, the tailor continues to manufacture these clothes. Strangely, the more bizarre and impractical his creations appear, the more he seems content and pleased with his work.

But if there are individuals who will manufacture anything, there are others who will find a use for their products — particularly if they are free. Occasionally, people with vision, those in desperate need, or those who are just plain greedy, rummage through the tailor's stockroom digging out garments which seem to fit their immediate purpose. When such a fit is found, the possible importance to humanity should not be underestimated.

Models created by the strange breed known as theoretical mathematicians allow us to perform operations which are absolutely prohibited in the physical world. It is not just a question of resources — something all the king's horses and men *could* do, given infinite time. Indeed, mathematics allows us to hastily perform tasks which could not be started at all with the physical manipulation of the objects involved.

In a real sense, theoretical mathematical models allow us to transcend limitations of time and space. Mathematicians have been living in the 4th, 5th, 6th, and 7th dimensions for hundreds of years. Today, as quickly as the computer flashes, we slide effortlessly and silently through a solid wall of the hardest steel, juggle particles so small that we cannot imagine them, or neatly pick up the planet Jupiter and move it a million miles from its orbit.

The mathematical models employed in psychological measurement do not allow for feats as exciting as those mentioned above. One extremely important formulation has already been mentioned: It is the *normal curve*, whose application in psychological testing was first introduced by Francis Galton.[5] The normal curve is one garment which has wide application. Like other mathematical creations, the normal curve does not exist on the shelf of your local grocery or hardware store; it was generated through calculus, and is entirely theoretical. The relevance of the normal curve to psychological testing will be discussed later.

Before continuing our discussion of the role of mathematics in psychological testing, a summary will most conspicuously allow the reader to compare his confusion with the writer's. (1) Numbers are a specialized language that probably evolved in much the same way as other verbal behavior. (2) At first, numbers represented objects in the physical world. Operations performed on those numbers (which were marks or scratches) corresponded in a direct manner to operations which could be performed on the objects themselves. Thus, the primary postulates of the natural number system were no more than relationships existing in the real world translated to the slate or paper. (3) Initially, these were relatively simple mathetical operations that could be performed on objects themselves. As objects became too numerous or difficult to handle, however, convenience was replaced by necessity. This marked the first instance in which mathematical operations transcended human capability. What the king's horses and men could not do, mathematics provided for handily. (4) Soon operations were made on numbers which the material and structure of things prohibited in the domain of objects. New systems were generated on the basis of postulates which often appeared contrary to real-world conditions. Historically, these formulations, regardless of the manner or purpose of their origin, have enabled us to solve problems which formerly thwarted all efforts at solution.

In returning to the topic of measurement and the language in which measurement statements occur, it can be seen that differences occurring in numerical languages, besides having an unrestricted range, lend themselves to all of the operations of mathematics. Measurement systems which result in statements expressed in non-numerical language are severely limited in terms of the manipulations which can be performed on them. It is a simple matter to add up statements of 43, 21, and 6, but how can one add up statements of "very good," superb," and "fantastic"? And, indeed, what is the square root of "wonderful" and the logarithm of "strongly agree"?

Psychological tests typically produce outcome statements in numerical language. These results are called test "scores". Test scores can be added, divided, subtracted, squared, or subjected to the full range of mathematical transformations. The practical advantage of mathematical treatment of test results, and its importance to psychological testing, will hopefully become more apparent in the following chapters.

FOOTNOTES

[1] Brumbaugh, R. S. Pythagoras and His School. *The Philosophies of Greece.* New York: Crowell, 1964, 30 — 42.

[2] Skinner, B. F. A Functional Analysis of Verbal Behavior. *Verbal Behavior.* New York: Appleton-Century-Crofts, 1957, 1 — 12.

[3] Skinner, B. F. *Verbal Behavior.* New York: Appleton-Century-Crofts, 1957, 469.

[4] *Ibid.*, pp. 461 — 470.

[5] Anastasi, A. Functions and Origins of Psychological Testing. *Psychological Testing.* New York: Macmillan, 1968, 7 — 8.

REFERENCES

Anastasi, A. *Psychological Testing.* New York: Macmillan, 1968, 7 — 8, 21 — 27.

Brumbaugh, R. S. *The Philosophies of Greece.* New York: Crowell, 1964, 30 — 42.

Freeman, F. *Theory and Practice of Psychological Testing.* New York: Holt, Rinehart & Winston, 1962, 5 — 7.

Guthrie, W. K. Pythagoras and the Pythagoreans. *A History of Greek Philosophy.* New York: Cambridge, 1962, 212 — 215.

Hilgard, E. & Bower, G. Mathematical Learning Theory. *Theories of Learning.* New York: Appleton-Century-Crofts, 1966, 334 — 338.

Kline, W. E., Oesterle, R. & Willson, L. M. *Foundations of Advanced Mathematics.* New York: American, 1959, 474 — 476.

CHAPTER 6

Descriptive Statistics

PRACTICAL MATHEMATICS

Among the mathematical advantages to which measurement statements expressed in numerical form lend themselves is statistics. Statistics is an applied form of mathematics having two subdivisions. These two subdivisions are *descriptive statistics* and *inferential statistics*. In this chapter, descriptive statistics will be considered.

When psychological tests are administered, there is almost invariably a need to deal with more than one test score or result. In some instances, a particular test is administered to the same individual in different situations or at different times. More commonly, a test will be applied to many individuals. Eventually, dozens, hundreds, and even thousands of measurements or test scores may result, and even when organized in neat rows and columns, these data may become unwieldy and difficult to interpret; they may simply not "make sense" to the reader. One writer has argued persuasively that humans cannot respond correctly to a particular stimulus configuration when more than eight separate stimulus elements must be considered simultaneously.[1]

Suppose you are confronted with a panel of lights, and below the lights are a number of buttons, each light having a corresponding button somewhere below it. Your task is to watch the lights and push the appropriate button as the light comes on. When only three lights and corresponding buttons are employed, the task is simple; as more buttons are added, it becomes increasingly difficult. You will find, unless you are super-human or discover some method of selectively disregarding some of the lights, that the time between light onset and button push becomes quite long when more than eight lights are employed. Furthermore, practice doesn't seem to help; the system, it would seem, reaches a saturation point. Columns and rows of data, regardless of how neatly and systematically they are organized, quickly become confusing for even the most experienced.

SMEARY ROCK STATISTICS

In order to give the topic some all-American relevance, let us assume that you are a football coach who is preparing for the big game of the season against the fantastic Smeary Rock eleven. The day before the contest, cruel fate strikes — your first-string quarterback, young Ferlin Fhagget, executes a tour jeté incorrectly in his ballet class and injures his leg so severely that even the slight pressure from his panty-hose makes him writhe in pain.

As coach, your task is clear — you must choose from among the three back-up quarterbacks; and it is no mean task, since you have had little opportunity to observe them under game conditions. In an effort to improve on a coin toss, some data are hastily gathered which were made during previous scrimmages. These data, given below in Table 1, show the passes completed by the three candidates along with the yardage gained. The data are incomplete, since they do not reveal how many incomplete passes were thrown, nor do they show how many quarters or how much time each of the records represents.

It is impossible to employ the data, as they are presented, to make an intelligent choice. Obviously, the performance of the three hopefuls must be compared, but first the data presented must be changed to a form that permits such comparison. In short, the data for each individual must be described in summary form. This brings us to the definition of a descriptive statistic.

A descriptive statistic is a statement, usually occurring in numerical form, which summarizes or describes an array or distribution of individual statements.

Individual outcome statements or measurements do not necessarily have to be in numerical language, and neither do statistics. When a group of individuals composing the membership of a club or fraternity are summarized or described as "cool", "gross", or "snobbish", such statements are, by our definition, statistics. Statistics of this sort are not generally helpful, however. Comparisons are not easily made among them and they do not lend themselves to mathematical treatment.

When a collection or array of individual statements or measurements are to be summarized, whether in natural or numerical language, there are many different aspects or dimensions which may be singled out for attention.

MEASURES OF CENTRAL TENDENCY

One aspect of a collection which may be used to describe it is the "central" or "middle" value of the collection. The highest or lowest value may also be emphasized, as when the strength of an army is characterized in terms of its strongest or most ferocious warrior. Strength can

TABLE 1.

Passes Completed by Three Reserve Quarterbacks

Bart		John		Craig	
Passes	Yards	Passes	Yards	Passes	Yards
1	0	1	0	1	90
2	52	2	25	2	0
3	0	3	24	3	0
4	0	4	0	4	0
5	22	5	34	5	47
6	11			6	0
7	2			7	2
				8	4
				9	1

also be described by considering the weakest and most timorous member of the army. Either one of these may be misleading, because a single member who is exceptionally competent or incompetent in relation to the majority of the members does not accurately represent the effectiveness of the entire group. An opposing general runs the risk of grossly overestimating or underestimating his enemy's potential if such descriptive statistics are employed. In this instance, an opposing general would do much better to use as a representative someone who was in the middle of the total distribution of strength measurements.

There are three descriptive statistics which are normally employed to represent the "middle" of a collection, array, or distribution of individual scores or measurements. These statistics are the *mean*, the *median*, and the *mode*. The three statistics are referred to as "*measures of central tendency*".

The measure of central tendency most commonly employed in psychological testing is the mean. The mean is simply the *arithmetic average* of a group of measures. It is derived by adding all the scores and dividing this sum by the total number of scores which were added.

The second most commonly employed measure of central tendency is the median. The median is that score or measure which exceeds in magnitude half of the scores in the collection, and is exceeded in magnitude by the remaining half of the scores. The median has more to do with ranking. It specifies the position of one score relative to the others in the array of scores.

The third measure of central tendency is the mode. The mode is that score measurement which occurs most frequently in the collection. The term mode is used in much the same sense that it is in everyday language. In fashion, for example, the mode refers to the most frequently worn style of clothing; this is exactly the same connotation that the statistical mode has.

Many students ask which of the three measures of central tendency mentioned above is the correct one to use. By now you can probably predict the author's answer to this one: "correctness" is not the pertinent consideration; the issue is rather which one *works* the best in a given situation.

Let us calculate these three measures of central tendency for the data presented in Table 1. In doing so, we can perhaps help the coach make the wisest selection.

The mean, you will remember, is calculated by adding up all the scores and then dividing this sum by the total number of scores in the distribution. When these calculations are performed, the following means are obtained.

Mean Yardage

Bart	John	Craig
12.4	16.6	16.0

If the mean is employed, John will get the nod to start in the big game against Smeary Rock, because his *average* yardage per completed pass is the highest. When the medians are determined for each candidate, John is also favored.

Median Yardage

Bart	John	Craig
2	24	1

The mode for each of the three men is zero; thus it is of no aid in selecting among the candidates.

Two of the three measures of central tendency favor John. Unless other contradictory data are available, the coach would probably do well to abide by this outcome.

Statisticians usually prefer the mean as the measure of central tendency because each and every measurement in the group of measurements contributes to this statistic. In contrast to the mean, the median is much less affected by the magnitude of specific scores. The scores are simply *ranked* in order of magnitude, and the middle score is designated as the median. Even a large change in the amount of a particular score will have either little, or no, effect on the median.

No mathematical procedures, not even ranking, are required to find the mode. A mere counting of each of the scores will reveal the mode.

Regardless of what some statisticians say, the measure of central tendency which should be used is basically a practical issue. Specific instances where one is superior to the others can always be found.

MEASURES OF VARIABILITY

Centrality, or middleness, is only one aspect of a collection of scores which may be described by a statistic. Among those most frequently employed in psychological testing, other than measures of central tendency, are statistics which summarize the *variability* occurring in the distribution of measures.

A psychological test, we have repeatedly stressed, is a difference-making machine. When a test is applied to many individuals, or applied to the same individual on repeated occasions, many scores or outcomes are generated, each of which will probably differ in some degree from the other. When scores are considered as a collection or distribution, the degree to which each score differs from the other, or from some reference point in the distribution, can be described. This feature of a collection of scores is called *variability*. Just as there are several statistics which represent central tendency, so there are several statistics which summarize the variability occurring in any array of scores or measures.

A very simple and sometimes useful statistic of variability is the *range*. The range is derived by relating the smallest score in the collection to the largest score. Often this is done by presenting the smallest score, separated by a dash from the largest score, in this manner: 25 - 98. In other instances, the smallest score is subtracted from the largest, giving the spread of points which separate the two.

The range for the passing yardage of the three quarterbacks discussed earlier are presented below.

Bart	John	Craig
0 - 52	0 - 34	0 - 90
52-0 = 52	34-0 = 34	90-0 = 90

Our coach would be wise to be concerned about variability in performance, since, other things being equal, he would prefer his quarterback to be consistent in his performance. However, when the range is employed as a statistic to express variability in the passing data, it is of little help to the coach. Since the lowest score for all quarterbacks is the same (0), the range gives no more information than the largest score, or, in this case, the longest distance covered by a completed pass. With many types of data, and for many purposes, the range is not a useful statistic.

Two more statistical descriptions of variability are the *variance* and its square root, the *standard deviation*. The variance, and therefore, the standard deviation, employs the mean as a *reference point*. An "average" amount by which the individual scores in the collection differ from the mean is expressed by these statistics. The variance is calculated by first obtaining the mean of the distribution and then subtracting each score from the mean. This procedure tells us how much each score *deviates* from the mean. Individual deviations are then squared, added together, and this sum is divided by the number of scores in the distribution. The standard formula given for the variance is:

$$\frac{\Sigma(\overline{X}-X_i)^2}{N}$$

Of course, X_i stands for each individual score; $\overline{X}$ for the mean; N for the number of scores in the distribution; and the funny Greek thing, Σ, is a summation sign, which means that, after being squared, the deviations are summed. Finally, the result of all these operations in the numerator is divided by N.

After the variance has been determined,

the standard deviation can be obtained by taking the square root of the variance. The full formula for the standard deviation simply involves placing a radical sign over the formula for the variance. (Note: When actually computing these statistics, say as part of some more complex statistical operation, other formulae, which are fully equivalent to those just presented, are employed for the sake of efficiency.)

Below, the variance and standard deviation are given for each of the three quarterbacks (the reader should perform the actual computations for himself at this point). This is the least we can do for the coach before he experiences his Saturday afternoon gridiron fiasco.

Bart
  Variance: 321:84
  Standard Deviation: 17.94

John
  Variance: 195.72
  Standard Deviation: 13.99

Craig
  Variance: 891.61
  Standard Deviation: 29.86

Inspection of the computational steps for the variance and standard deviation reveals that these statistics are types of two or more different distributions.

In the next chapter, it will be shown that the mean, variance and standard deviation are basic components of the second type of statistics, which are called *inferential statistics.*

In the meantime, the writer has amply demonstrated to his own satisfaction how difficult it is to make the mean, median, summary statements of the variability among all the scores, with the mean as a reference point. The standard deviation is a unit of variability in much the same way that a pound is a unit of distention of the spring or deflection of a balance bar. As a descriptive statistic, the standard deviation is more widely employed than the variance, since it permits immediate, standardized comparisons of the variability of mode, range, variance, and standard deviation really "live". Certainly these concepts are not the kind that can launch a revolution or excite great throngs of people to the point of a standing ovation. The student may skeptically ask for some demonstration of the magical properties mentioned earlier. Be patient, if not merciful, and read on.

FOOTNOTE

[1] Miller, G. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review.* 1956. 63, 2, 81 — 97.

REFERENCES

Anastasi, A. *Psychological Testing.* New York: Macmillan, 1968. 40 — 45.

Burke, C. Additive Scales and Statistics. *Psychological Review.* 1953. 60, 1, 73 — 75.

Freeman, F. *Theory and Practice of Psychological Testing.* New York: Holt, Rinehart & Winston, 1962. 24 — 33.

Games, P.A. & Klare, G.R. *Elementary Statistics: Data Analysis for the Behavioral Sciences.* New York: McGraw-Hill, 1967, 147 — 149.

McNemar, Q. *Psychological Statictics.* New York: Wiley, 1969. 14 — 25.

Miller, G. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review.* 1956. 63, 2, 81 — 97.

Stevens, S.S. On the Averaging of Data. *Science.* 1955. 121, 113 — 116.

CHAPTER 7

Inference, Samples, and the Normal Curve

YERKES STRIKES AGAIN

Let us begin this chapter with another story of unrequited love. Once again, one of the main characters of the piece is the noted animal psychologist, Robert Yerkes. You will remember Professor Yerkes as the same individual who distinguished himself in chapter three by boiling a love-smitten frog. His appetite undoubtedly whetted, Professor Yerkes undertook an entirely new form of torture. The victim on this occasion was a female frog named Eunice.

From the very beginning it was clear that Eunice was hopelessly in love with Yerkes. It is difficult to pinpoint the origin of the fond regard which grew into boundless passion. Unquestionably, the countless gifts he bestowed on Eunice contributed to the process. For instance, every day he would bring her several of the choicest delicacies — giant cockroaches. Next to Dr. Yerkes, there was nothing Eunice liked more than a juicy cockroach.

But there were other reasons Eunice was led to believe she was something very special to the dashing Yerkes. He had constructed a pedestal of shiny metal for her to sit on. Each morning Dr. Yerkes would pick up Eunice, place her on the pedestal and present her with a choice morsel of cockroach. You can now understand Eunice's feelings toward the psychologist. What girl, even if she is a bull frog, can resist a man who places her on such a pedestal?

But happiness for romantic frogs, it would seem, is, at best, a fleeting thing. One black morning when Dr. Yerkes came to get Eunice, she sensed a change in his demeanor. There was a certain air of objectivity about him which made her cold blood run even colder. Had Eunice not been so busy attending to Dr. Yerkes, she would have noticed a wire running from an electric shock source to her pedestal.

As always, the customary cockroach was presented, and Eunice's heart jumped with joy. It is said that love is blind, and in this case it was, because Eunice failed to notice a second wire running from the cockroach to the other pole of the shock source. With complete confidence she extended her tongue and wrapped it around the cockroach. Instantly, a jolt of electricity shot through her body. Wracked with spasms, she leaped off the pedestal, convulsing pathetically.

There is a strong tendency to speculate on Eunice's feelings after her shocking ordeal. We shall resist that tendency. Let it merely be recorded that from that day on Eunice refused to eat. Time after time — doubtlessly filled with remorse — Dr. Yerkes appeared offering even larger and more delectable cockroaches than he had in the past. But Eunice would not eat. Slowly, she lost weight. Her muscles, once taut and vibrant, became flaccid and unresponsive. One morning Dr. Yerkes found Eunice dead.

Below is Dr. Yerkes' published statement on Eunice's passing to the eternal lily pond:[1]

> "While making measurements of the frog's reaction time to electrical stimulation, I noticed that after a few repetitions of a 2 volt .0001 ampere stimulus an animal would make a very peculiar noise. The sound is a prolonged scream, like that of a child, made by opening the mouth widely . . . The question arises, is this scream indicative of pain? . . .
>
> "Are we to say that the weak stimulus is painful because of increased irritability, or may it be concluded that the reflex is, in this case, like a wink or a leg jerk, or the head lowering and puffing, simply a forced movement, which is to be explained as an hereditary protection device, but not as necessarily indicative of any sort of feeling.
>
> "Clearly, if we take this stand, it may at once be said there is no reason to believe the scream indicative of pain at any time. And it seems not improbable that this is nearer the truth than one who hears the scream for the first time is likely to think."

Yerkes' experiment is relevant to human behavior. The instance demonstrated (albeit at Eunice's expense) is one referred to as "traumatic avoidance". It occurs when a previously preferred or at least neutral object or situation is abruptly paired with an aversive stimulus. As a result of this pairing the individual will do all he can to avoid contact with or even proximity to the object he had earlier regarded as pleasant or harmless.

Whether the behavior be of frog or human, the interesting feature of traumatic avoidance is that it is "irrational". It persists even after massive evidence has been presented to show that the association of the aversive stimulus and the "feared" object was a one-shot thing, possibly never to occur again.

In Eunice's case it appeared that after one single aversive episode with a cockroach, a hasty "conclusion" was made that all subsequent episodes with cockroaches would also be aversive.

In most cases we would be wise to follow this sage advice of our elders: "Don't judge the barrel by one rotten apple", or, in the words of great-grandmother Wechsler, "One swallow does not a summer make."

From this standpoint we may say that once shocked, Eunice literally jumped to conclusions. To assume that if one cockroach accorded her such bad treatment, all subsequent cockroaches would perform in the same manner, appears to be faulty reasoning.

Of course, it can be pointed out that, being a frog, Eunice was incapable of reason, thought, consideration, or any of those behaviors man engages in when he is being rational. While this may serve to vindicate Eunice, it places man, whose greatest boast is his logic and rationality, in a bad light. The fact is that humans exhibit traumatic avoidance which is not substantially different from that exhibited by Eunice. Fortunately, if this irrational avoidance behavior causes the individual extreme difficulty, a few visits to the therapist can generally alleviate the problem. Thus far, however, there are no published reports indicating that the treatment has been extended to frogs.

Humans not only jump to conclusions by being unduly influenced by a single aversive instance. The reverse reaction is even more prevalent. They often fail to jump when they should. Mild discomfort may correlate with steadily worsening physical conditions leading to severe illness or death. Cigarette smokers, for example, continue to smoke year after year although they exhibit a chronic cough or have persistent difficulty in breathing. Many others continue to eat highly seasoned foods which commonly cause digestive upset and discomfort.

Thus, in some instances, just as Eunice, we learn not too wisely, but too well. In other instances, even in the face of steadily worsening conditions, we learn not at all.

INDUCTIVE SCIENCE

Surely, you may suggest, man can do better than the frog. There must be a way in which he can use his past experiences more intelligently to benefit from them now and in the future.

Fortunately, you are correct. Through his ability to symbolize past experience in language, the history of an individual, or even a civilization, can be analyzed in an objective fashion; and based on this analysis, effective courses of action can be taken.

As a general practice, this method can be called *inductive science*. The process in which the symbolized past is employed to predict the future is called *inference*. Your local weatherman employs inference in making his predictions. Let's look further into the possible value of inference.

The previous chapter — as you will remember — mentioned two subdivisions of statistics. The first, descriptive statistics, we have already discussed. The second involves the use of statistics to make inferences from past observations to the future, and is called *inferential statistics*. The use of inferential statistics enables us to make predictions that are far more accurate and specific in nature than those based on intuitive guesses or common sense.

All inference, including statistical inference, operates on the principle that we may learn what to expect from a large collection of objects, things, or experiences by examining the properties of a sample taken from that collection.

Statistical inference is a method employing mathematics which not only allows us to generalize from a sample to a larger collection, but also yields some statement of the likelihood that predictions based on the sample will be correct.

POPULATION AND SAMPLE

In statistical inference a total collection of objects, things, or experiences is called a *population*. Any part less than the whole of a population is, of course, a *sample*. These two terms, population and sample, are crucial to statistical inference.

Students often become confused about the relationship between a sample and a population. First of all, a population is a defined quantity or entity. A population of men could include as few as one man, provided the population was defined as, say, all individuals named Phillip Jones, living in Kansas City, Mo., at 2203 Maple, being 41 years old, having a wife named Patricia, etc. In this example, a sample of the population containing at least one person would be exactly the same as the population, and inferential statistics would not be needed. In all other cases a sample contains fewer objects, things, or observations than exist in the population. Regardless of how large the population is, and how nearly the same in size the sample is, a sample is still a sample, provided it contains fewer members than the population. Thus a collection of 5,000,000 minus 1 is still a sample of a population containing 5,000,000 objects or observations.

Another source of confusion revolving around the concepts of population and sample has to do with the nature of the composition of a population. A population need not be people. It can be balls, numbers, sparks, doors, chairs, populations, or anything defined as the totality of a class of things under consideration.

A population may be finite, as beans in a jar, theoretically finite, as all the mosquitoes that will ever exist, or infinite, such as time.

Often a sample is a portion of a population of objects which exists at a given point in time. A handful of tickets selected from a shoe box or a sample of residents in a particular geographical area are examples. In other instances, a sample comprises observations of events or phenomena which will be employed to make predictions of future events or phenomena.

In psychological testing both types of sample-population relationships are important. We may observe the abilities of a sample of students taken from a particular university, and from this sample make inferences about the abilities of the entire student body population. Thus, both sample and population currently exist.

For a psychological test to be useful, it must have relevance to future events or situations, so that the abilities of an individual may be assessed today and predictions, regarding future performance in college or on the job, made.

STATISTICAL INFERENCE

In order to demonstrate how statistical inference works, let us consider a realistic problem, one which might be encountered on any college campus.

Imagine that we were looking for a date for our best friend who, although a stupendous track star, is only 5 feet tall. Knowing the average or *mean* height of the girls in the dorm from which a blind date is to be chosen would help us in selecting a suitable girl. Practically any blind date coming from Amazonia Dorm, where the average height is 7'2", would in all likelihood send our friend scurrying for a footstool when the time came for the goodnight kiss.

From the previous chapter we learned that some information regarding the variability of heights would also be helpful. The *range* of heights would help us greatly, and, of course, the *variance* or *standard deviation* would be of even greater utility. (Remember, the standard deviation is the square root of the variance.) We might be better off to select a girl from a dorm with an average height of 4'8" and little variability than to take a chance on a dorm where the average height is only 4'5", but the variability is great.

Actually, the most information would be available (short of knowing who the girl will be) if we could have a listing of all the heights of girls in a particular dormitory. We could break down the list into height categories such as 4'0" — 4'2" — 4'3" — 4'5", etc. By counting the exact number of girls in each category, we could compute the exact probability of obtaining a girl within the acceptable height range. By having such a listing for each of the dorms, you could select that dorm which is most likely to yield an acceptable blind date. Considering any girl taller than 5'0" as being unacceptable to our friend, and turning to Diminutive Dorm, we find that only 10% of the girls residing there are taller than 5'0". Thus, the probability of obtaining a girl who is unacceptable is only 1 in 10, odds which our friend may be willing to risk.

In the example above, girls' heights were placed into equal-sized categories, called *class intervals*, which were then ranked in order of magnitude. The number of girls in each category was also recorded, yielding a collection of measurements technically called a *frequency distribution*. Obviously, knowing the manner in which heights are distributed in the population gives us considerable information on what to expect of a sample taken from that population. In this example, arranging the heights of all the girls in each dormitory into separate frequency distributions would be too much work, even for a best friend. Surely there must be a less laborious method to reduce the uncertainty of the blind date.

THE NORMAL CURVE

Now is the precise moment to introduce the *normal curve*, which has received some advance publicity in preceding chapters. The normal curve is a mathematical formulation which was generated by calculus, and languished on the coat racks of the mad mathematical tailor before Francis Galton[2] saw it would fit human structural and behavioral characteristics.

Most observations, when plotted or arranged in a distribution, assume the same general shape as the normal curve. Heights, weights, most test scores, blood counts, freckles, nose lengths, number of hairs on heads, to name only a few, are all things which would, when placed in a distribution, assume the shape of the normal curve.

The proportions of the normal curve are its most important features. Just as girls with great and absolute differences in size, height, and weight may have the same pleasing hour-glass figure, many different distributions with large discrepancies in means and standard deviations may all assume the form of the standard normal curve. Once we know that a group of observations is distributed in a normal curve, we can, without bothering to arrange the observations in a distribution, estimate the likelihood of obtaining a score of a particular value. As a matter of fact, if we know that a population is normally distributed, we can estimate the probability of coming up with any particular value by the expedient of taking a random sample from that population and applying the laws of statistical inference. But we're getting ahead of our story.

In our example we were looking for a girl of moderate height to go out with our track star. If we had the means and standard deviations of heights of girls for each of the dorms on campus, and if we were convinced they were normally distributed (the means and standard deviations, that is) within each of the dorms, we could tell immediately the risk of picking a girl taller than our friend, and we could do so without bothering to look at a list of heights. Below in Fig. 1 is the standard normal curve.

The height of the curve gives information on the relative frequency to be expected at any point along the horizontal axis, which is technically called the abscissa. In a normally distributed collection. the most frequently observed values will be those near or around the mean. This is a way of saying that most people are average. As values deviate from the mean in either direction, they become less frequent. Since the normal curve above is generalized, expectations are not stated in frequencies, but in percentages. As we move to scores further from the mean, the percentage of cases decreases.

Now let us take this generalized curve and apply it to the problem of selecting a date for our friend. If at all possible, we would like to get our friend a date with a girl from Horni Dorm, since girls there are noted for their intellectual prowess, and we would like to insure that our friend will have a stimulating evening. We find that the mean height there is 5'0", and the standard deviation is 2.0". We know also that heights in the dorm are normally distributed. These specific values can now be placed in the generalized form of the normal curve. This is depicted in Figure 2.

Immediately, we can see that half of the girls in Horni Dorm are going to be taller than 5'0", taller than our friend. However, by consulting the curve, we see that of the 50% of the girls taller than he, approximately 34% exceed the mean of 5'0" by *less than* one standard deviation of 2.0". Thus, of the girls taller than our friend, only approximately 16% exceed his height by more than 2.0" (50% minus 34%). This means that if those 34% between the mean and +1 standard deviation can stoop a little, 84% of the girls from Horni Dorm could be acceptable (50% less than the mean of 5'0" *plus* 34% between the mean and +1 standard deviation equals 84%). Consequently, we decide to take our chances on Horni Dorm.

THE NORMAL CURVE AND PSYCHOLOGICAL TESTS

In previous chapters we have referred to measurement as a procedure for creating differences. When a psychological test is applied to a group of individuals, a particular score is generated for each individual. We would expect that only rarely would two people receive exactly the same score. You will recall from a previous chapter that Sammy Slick's test was rejected by us because all subjects received an identical score of 437. It was made clear that the test which generates the largest number of differences has the greatest chance of being successful.

The fact that the differences created — when different girls subjected to a ruler fell into a distribution assuming normal proportions — allowed us to compute with precision the probability that a blind date would be of suitable height. Differences in measurements generated by a *psychological test* can be treated in the same manner as physical differences. Once we know the mean and standard deviation of such a distribution, and are assured it is *normally* distributed, we can predict quite precisely the probability that the score of an individual selected at random from that distribution will be above or below a particular value.

But some may dismiss this apparent legerdemain as an intellectual confidence game. True, once the mean and standard deviation of a distribution are known, certain predictions are possible. It should be remembered, however, that each and every score in the distribution must be available before the mean and standard deviation can be computed. Presumably, the same predictions are possible through inspec-

FIGURE 1

Generalized Normal Curve

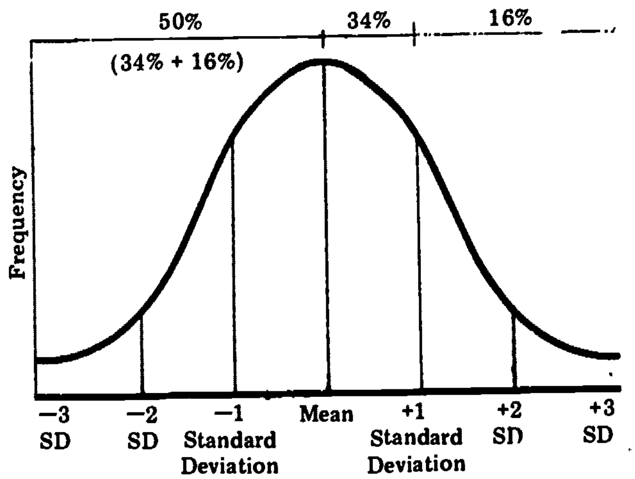
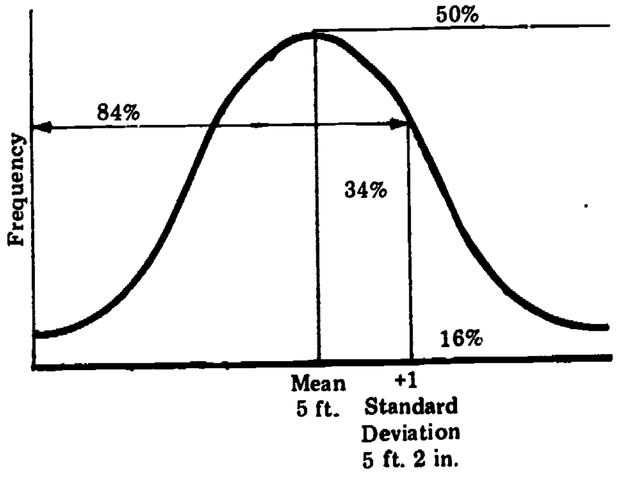


FIGURE 2

Generalized Normal Curve for Height of Dorm Girls



ting and counting scores as they appear in columns and rows, a method which would have the advantage of not requiring the computation of mean and standard deviation; but this method becomes very cumbersome as scores proliferate. A technique employing the normal curve, mean, and standard deviation, would perhaps be more convenient.

The importance of the normal curve is far greater than the simple convenience mentioned above; specifically, it allows us to move swiftly and precisely from a limited space and time which we have studied thoroughly, to a larger and perhaps unlimited space and time which we have not yet studied and, indeed, may never experience. We are, of course, referring again to the process of *statistical inference*, and we shall now delve into its mysterious properties.

LAWRENCE WELK AND POLITICS

Suppose as a part of a course assignment you interview one-hundred students regarding their political beliefs. When your results are compiled, all of the respondents agree on one thing — they will strongly support any candidate who resembles that virtuoso of the polka, the head bubble himself, Lawrence Welk.

Hearing quite by accident of your survey, one political hopeful dedicated, above all, to winning elections, hires a plastic surgeon, a crack accordion teacher, and a Bavarian diction coach. His plan, he explains to you, is to campaign on a platform of free sauerbraten and mandatory accordion instruction for all citizens. One question has occurred to him, however. Can he be sure that the affinity for Lawrence Welk manifested by the students in your survey will also hold for those outside your sample? In short, what the politician wishes to know is: can statements or conclusions made from experience with your limited sample be generalized to the entire student population?

Before you can answer, some preliminary considerations are in order. In the first place, a population is a defined entity. While the group of students in your survey can be considered a sample from a population of the totality of students in the world, such a comprehensive population is not necessary or practical. The politician is not interested in all students in the world; he is only interested in those who are eligible to vote in the particular election in which he is involved. The population which is relevant is that delimited by the voting requirements in the particular election in question. Presumably, only students meeting those requirements can possibly influence the outcome of the election, and it is they whom we must consider to constitute the defined population.

Assuming that the students interviewed are actually a sample from the population (i.e., they qualify as potential voters in the political contest) we are now ready to deal with the question posed by our politician: Can conclusions drawn from the sample be applied to the population with a reasonable degree of confidence?

The answer involves only common sense. Insofar as the sample is similar to the population in all relevant dimensions, statements true of the sample will also be true of the population. This answer, unfortunately, like most other common-sense answers, seems only to create more questions. Now, we must go on to ask the conditions under which we are justified in assuming that a sample will have features common to that of the entire population. Naturally, we could examine the population and compare it to the sample; but then, of course, there would be no utility in employing a sample.

KNOWLEDGE OF THE DISTRIBUTION

Often we have some prior knowledge of how a particular set of features or characteristics are distributed throughout the population. Red ink or dye added to a body of water will usually immediately dissolve and form a homogeneous pink solution. The process may be speeded by agitating the water. Once we know the typical manner in which the dye disperses itself uniformly throughout the water, we would not hesitate in judging the concentration of dye in the entire body of water from the analysis of perhaps just a few drops of the solution.

Other substances, such as pebbles, do not behave in a like manner. The distribution of pebbles in water will depend upon a variety of factors, including currents, the weight of the pebbles, and other physical characteristics. Hundreds of samples could be taken from all but the very bottom portion of the water and, almost certainly, none would contain pebbles. Samples taken in this fashion would lead you to believe that there were *no* pebbles in the water.

Perhaps this is the most opportune time to introduce a term having to do with the correspondence of a sample to the population from which it was drawn. When what is true of a sample is also true of the population from which it was drawn, the sample is said to be *representative*. Being necessarily smaller than the population, samples cannot be expected to contain as many dye particles or pebbles as the population. Representativeness would mean, however, that the *proportion* or *percentage* of dye particles or pebbles in the sample, and the population, are identical. Thus, if a sample is representative of the population, we can confidently judge what to expect in the population on the basis of our experience with the sample.

Returning to the problem at hand, can we be assured that the sample of 100 students with the fetish for Lawrence Welk is representative of the student constituency of the politician? The answer is that we probably cannot. In all likelihood the sample is not representative. The politician would do well to investigate further before carrying out the planned metamorphosis.

Unlike red ink or dye, human beings do not generally distribute themselves evenly throughout a geographical area. A sample of one-hundred students taken from one location cannot be assumed to be the same as a sample taken from another location. Polling the first hundred students one encounters shows commendable dedication to efficiency, but it will, undoubtedly, result in an unrepresentative sample.

Actually, the question of representativeness is much trickier than we've indicated, because even when you've taken all the recommended steps and precautions in the selection of a sample, you cannot be certain that it is representative of the population. Perhaps you've failed to consider some important characteristic in your sampling scheme; or else you may have assigned too much weight to one or more of the factors you deemed important. Such difficulties, however, do not mean that all attempts to secure representative samples should be abandoned. In many cases you can do no better than to try to assemble the most representative sample you can based upon your knowledge of the relationships between various population characteristics and the characteristic you're measuring. Obviously, the more you know about these relationships, the greater the likelihood that you will be able to generalize the findings from your sample to the population as a whole.

RANDOM SAMPLING

The uncertainties and difficulties associated with representative sampling can often be circumvented by adopting a quite different technique called *random sampling*. A random selection procedure is one in which *each member of the population has an equal chance of being chosen for the sample*. Practically speaking, there are numerous methods for selecting a random sample. Names corresponding to each individual in the population could be placed on slips of paper, which are then placed in a barrel. The barrel is then shaken or rotated, insuring that the bits of paper are thoroughly mixed. A sample of 100 slips can then be drawn, each slip having an equal chance of being selected. Such a sample is said to be random.

A random sample is not necessarily a representative sample; it is not likely to be closer to the truth than samples selected according to different procedures. Indeed, the major virtues of random sampling have little to do with the concept of representativeness. Random sampling is a powerful technique because *random samples behave in a predictable manner in accordance with precise mathematical laws.* Randomization enables us to associate *mathematical probabilities* with various events. Technically, it permits us to establish a confidence interval — the likelihood that the true population value is within a particular range on either side of our sample value. But more of this later.

There is another sampling procedure which combines random sampling and representative sampling. This procedure is called *stratified random sampling,* and a sample selected by this method is called a *stratified random sample.* Since random samples are not necessarily representative, there is always a degree of "error" associated with their use. A procedure which increased the representativeness of a random sample would also reduce the "error" — that is, it would increase the probability that the sample included essential characteristics in the same proportions as the population.

Such a procedure demands that something be known about the constituency of the population. For example, knowing the ratio of men to women, the distribution of socioeconomic factors, political affiliations, amount and type of education, and other factors may help to construct quite a representative sample. Suppose that in consulting his budget, a test constructor sees that he can only afford to employ 100 individuals in his sample. Knowing from previously collected data that 50% of the population is male and 50% female, that 25% are farmers, 25% factory workers, 15% college students, and that 75% are under 50 years of age, etc., the researcher will be in a position to assemble his sample according to the same proportions.

Respecting the proportions above, it can be seen that in a sample of 100 which the researcher can afford, 25 subjects must be over 50 years of age, 50 subjects must be males, 25 must be farmers, etc. Once the proportions are determined all individuals available who fit each category are included in a group from which the required number is drawn randomly. For instance, 25 subjects must be over 50 years of age. Initially, all individuals fitting this age requirement will be considered, and within this category names will be selected by a random procedure. Since this type of sampling involves grouping, or stratifying, subjects according to certain known characteristics of the population, and since it also involves random selection within the known boundaries, the term *stratified random sample* is used.

In stratifying a sample, the test constructor takes into consideration those known attributes of the population which, if not represented in the sample, would tend to make conclusions drawn from it unrepresentative of the population. To a biochemist who wished to make a statement about the typical red blood of Americans, such things as age, sex, diet, and general health might all be relevant factors. It is not as apparent that traits like religious preference, I.Q., and marital status, although perhaps readily available from the data at hand, would be of any importance in selecting his sample. Such variables would not be considered in the stratification.

SAMPLE SIZE AND REPRESENTATIVENESS

Before we go off the deep end with discussions of sophisticated sampling strategies aimed at insuring representativeness, one obvious relationship must be considered; namely, the fact that the *size* of the sample has a great deal to do with the possibility that it will be representative of the population. Independently of how the sample is selected, as its size increases to the point that it contains almost as many elements or members as the population, it becomes more and more representative. All things being equal, the larger the sample size relative to the magnitude of the population, the more representative it will be.

In many cases, however, the population may be so large as to be infinite, or at least practically so. It would be impossible for even the most loquacious and ambitious interviewer to speak to enough people to improve significantly on the relative size relationship of sample to population. Clearly, the sample could never begin to approximate the magnitude of the population.

But knowledgeable and ingenious men need not be intimidated, even by the awesome specter of infinity. With the aid of inferential statistics, quite accurate predictions or inferences are typically made to populations of theoretically infinite magnitude from samples of 100, 50, or even less. Furthermore, definite and exact statements can be made regarding the probability that such predictions are correct.

Let us be so presumptuous as to demonstrate the mathematician's machinations. First, however, you must be forewarned of one condition which must be satisfied before the mathematical laws of probability can be applied to statistical inference. Samples must be selected randomly from the population — or else!

For the moment, let us assume that all samples are selected by a random procedure. While we're at it, let's change history and assume that the 100 students exhibiting the extreme penchant for Lawrence Welk were not an unrepresentative sample, as indicated earlier, but were selected randomly from the population. We must also qualify the manner in which the outcomes of interviews are expressed. Rather than a single gasp of undying admiration for Lawrence Welk, assume that devotion was expressed in the form of a score which increased in magnitude as the degree of affection increased. Under these conditions, scores taken from the sample of student respondents can be summed up and a mean computed. Since our sample is the only information we have about the population, the best guess we can make about the mean of the population is to say that it is the same as the sample mean.

If many random samples of 100 students were drawn from the population, we would expect such samples to be comprised of different individuals, although some persons might be selected for more than one sample. Since individual scores are variable, and, as just mentioned, each sample is likely to contain a different combination of individuals, means computed for each sample will differ. Nonetheless, at any given time, the best estimator of a population characteristic is that characteristic observed in the sample. Incidentally, when a descriptive statistic such as a mean, median, or standard deviation is computed for an entire population, it is called a *population parameter* or just a *parameter.* In fact, any characteristic of the population is traditionally referred to as a parameter. Similar characteristics computed for samples are called *statistics.*

In the problem under consideration, a politician wishes to know if data collected on a sample of student constituents will also be observed in the population. Will the mean computed for the sample of students who evinced a fondness for Lawrence Welk be observed if all the students in the population are interviewed and a mean computed on those scores? Again, common sense tells us that a statistic (e.g. the mean) derived from any sample, regardless of the size, will be likely to differ somewhat from the corresponding population parameter. Common sense does not say, however, how much a sample mean is likely to differ from the population mean. Neither does it tell us if increasing the sample size will improve the accuracy of predictions made from a sample to the population. (It is advisable to remember that "common sense" once led men to conclude that the earth was flat.)

STANDARD ERROR OF THE MEAN

In the last chapter, the introduction of the standard deviation was accompanied by great fanfare. Perhaps at that time you wondered why all of the hullabaloo for something that was, after all, only standard. In any event, you will remember that the standard deviation was a descriptive statistic related to the variability occurring in an array of scores. It was computed by subtracting each score from the mean, squaring those differences (deviations), adding the product, dividing by the number of scores (n), and finally extracting the square root of the resulting mess. More succinctly,

$$\text{Standard Deviation} = \sqrt{\frac{\Sigma(\bar{X}-x_i)^2}{n}}$$

where X_i indicates individual scores, $\bar{X}$, the mean, and n, the number of scores. Of course, the funny thing (Σ) is Greek to us all (even to the Greeks). It is called Sigma, and means that all the squared deviations are added together. In English, Σ is aptly called a *summation sign*.

There is another formulation which has almost the same form as the standard deviation. Like the standard deviation, it is also a measure of variability. Let us visualize a population from which a large number of random samples of the same size are drawn. If, further, a mean is computed for each sample, we would expect those means to differ from each other, because the composition of each sample is different. The computation of variability among the means of samples begins by comparing each of them to the mean of the population. Thus:

$$\sqrt{\frac{\Sigma(M-\bar{x}_o)^2}{N_s}}$$

Now compare these two formulae directly, paying particular attention o the differences in the subscripts. In the latter formula, $\bar{X}$ stands for individual sample means, M for the population mean, and N_s for the total number of samples drawn.

This measure of variability describes the way in which the means of random samples deviate from the mean of the entire population. Whether the number of samples drawn is 2, 25, or even 100, this formulation is not altogether inspiring if it is only another descriptive statistic, a standard deviation of sorts. But let us not prematurely dismiss this formulation. Statisticians have demonstrated that if a theoretically large number of samples are drawn from a population, and if the means of each sample are compared to the mean of the population, the formula above approaches the standard deviation of the population divided by the square root of the number of samples. Thus, when N is, theoretically, extremely large,

$$\sqrt{\frac{\Sigma(M-\bar{x}_o)^2}{N_s}} = \sqrt{\frac{\frac{\Sigma(M-x_p)^2}{N}}{n}} \text{ or } \frac{\sigma}{\sqrt{n}}$$

Before becoming confused, let us compare and delineate the three foundations we are currently considering. The standard deviation of a sample as introduced in Chapter 5 is expressed as follows:

$$s = \sqrt{\frac{\Sigma(\bar{X}-X_s)^2}{n}}$$

Notice that each individual score in the sample (X_s) is subtracted from the sample mean ($\bar{X}$). The denominator (n) represents the number of scores in the sample. The standard deviation of the sample is signified by the lower case "s".

The standard deviation of all the scores in a population is upon inspection almost identical to that presented above:

$$= \sqrt{\frac{\Sigma(M-x_p)^2}{N}}$$

In this formulation, however, notice that each and every score in the population (x_p) is subtracted from the population mean (M), which is likewise computed from all scores in the population. The denominator (N) is the number of scores in the entire population. The standard deviation of the population is, by convention, represented as σ, the lower case Sigma of the Greek alphabet. Thus *s* indicates the standard deviation of a sample and σ indicates the standard deviation of the population.

The third formula remains as yet unnamed:

$$\sqrt{\frac{\Sigma(M-\bar{x}_o)^2}{N_s}}$$

Here the means of individual samples ($\bar{x}_o$) are subtracted from the population mean (M). The denominator (N_s) refers to the number of sample means ($\bar{x}_o$) under consideration. As sample after sample is drawn, and the number of means (N_s) becomes very large, statisticians insist that:

$$\sqrt{\frac{\Sigma(M-\bar{x}_o)^2}{N_s}} = \sqrt{\frac{\frac{\Sigma(M-x_p)^2}{N}}{n}}$$

Since we have already identified the numerator of the latter formula as the standard deviation of the population (σ), we can express the formulation as:

$$\frac{\sigma}{\sqrt{n}}$$

Notice that the denominator ($\sqrt{n}$) refers to the square root of the number of scores in a sample.

This formulation, $\frac{\sigma}{\sqrt{n}}$, is called the *standard error of the mean.* It expresses the amount of difference we would expect to find when a mean computed from a sample ($\bar{x}$) is compared to the population mean (M). Practically speaking, this statistic allows us to estimate how far we are likely to be in error if we accept a mean computed from a sample as being representative of the population mean.

If the reader will recall, the purpose of this entire discussion was to point up the relationship between sample size representativeness. The *standard error of the mean,*

$$\frac{\sigma}{\sqrt{n}}$$

defines this relationship quite specifically. The standard error of the mean decreases (and thus representativeness increases) as the sample size (and hence, its square root) increases.

Suppose we are dealing with a population whose standard deviation is known to us. For illustration, let's use girls' heights as a somewhat familiar case in point. Knowing that the population standard deviation, σ, is 3.0 inches, it is possible to see specifically how drawing samples of various sizes will affect the standard error.

If a sample of four is selected, then:

$$\text{Standard error} = \frac{\sigma}{\sqrt{n}} = \frac{3.0\text{ in.}}{\sqrt{4}} = \frac{3.0}{2} = 1.45\text{ in.}$$

A sample of 16 decreases the standard greatly:

$$\text{Standard error} = \frac{\sigma}{\sqrt{n}} = \frac{3.0}{\sqrt{16}} = \frac{3.0}{4} = .75\text{ in.}$$

Standard errors have been computed for various sample sizes, given a standard deviation of 3.0 inches, and are presented below in Table I.

TABLE I

Sample size	4	16	25	100	200
Standard error	1.45 *inches*	.75	.60	.30	.21

It can be seen from the table above that increasing the sample size greatly reduces the standard error of the mean. However, beyond a certain magnitude, further increases cease to effect worthwhile reductions in the standard error of the mean. For instance, increasing the sample size from 4 to 100 decreased the standard error from 1.45 to .30 inches, a reduction of 1.15 inches. Adding another 100 units to the sample size, and increasing it to 200, reduces the standard error from .30 to .21, an advantage of only .08 inches.

This relationship between sample size and representativeness of sample to population has at least two important ramifications. On the practical side, the relationship tells the researcher or test constructor that beyond a certain point there is no great advantage to be gained, in representativeness, by increasing sample size. If, for instance, a herpetologist is measuring the caressability of king cobras, graduate assistants helping out in the research will be pleased to know that a sample of 200 or even 500 is not appreciably more representative than a sample of 100 or even 60.

The theoretical consideration stemming from this relationship is by far (although we could not convince the graduate assistants) more important. The fact that the standard error of the mean decreases as the absolute size, and not the relative size, of the sample increases means that a sample of modest size is appropriate for prediction to populations of infinite size. If representativeness were strictly dependent on the size ratio of sample to population, prediction to infinite or practically infinite populations would be impossible, because no researcher could ever manage such a large sample.

To summarize, the standard error of the mean represents a kind of average difference we could expect between means computed for samples and the mean of the population from which those samples were drawn. In this sense it indicates how close estimates based on a sample are likely to be true of the population.

Now, let's return to the example we used to introduce the topic of representativeness. Originally, you'll recall, we hoped that inferential statistics would help us solve a practical problem. A sample of 100 students were interviewed and responses converted to numerical scores. From these scores a mean was computed, which revealed that the students would overwhelmingly favor a political candidate who resembled Lawrence Welk in both manner and appearance.

But can we conclude that what is true of the sample will be true of the population? Ah ha! This looks like a job for the standard error! But first it must be computed, and this requires the standard deviation of the population (σ). Let us assume that our political friend obtains this information. In this instance, $\sigma = 100$.

The question we are asking is how much the sample mean ($\overline{x}$) can be expected to differ from the population mean (M). Of course, the sample mean can be computed directly. In this case, let us assume that it was found to be 75. The population mean (M) is not known. If it were available, the two means could be computed directly.

Computing the standard error:

$$\text{Standard error} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{100}} = \frac{100}{10} = 10.$$

Therefore the standard error of the mean = 10.

After computing the standard error of the mean, there is yet more work to be done. It is, you will remember, a type of standard deviation expressing the variability one might expect from means computed on sample after sample drawn from the population. Now it has been shown mathematically that the means of a very large number of random samples tend to assume a *normal distribution* — even if the original population is not normally distributed![2] This allows us to use the normal curve for our predictions. It is presented in Figure 3.

In the example above, a mean of 75 was computed from the sample of 100 students. Obviously the population mean could be either the same, larger, or smaller. In order to estimate the probabilities associated with these outcomes, we take our sample mean of 75, along with the standard error of 10, and refer directly to the percentages associated with the normal curve.

The probability that a mean, computed from a random sample of 100, will differ from the population mean by a given amount, or more, can be calculated from the standard error of the mean. The normal curve indicates that approximately 68 percent of all means computed from such samples would differ from the population by no more than plus or minus one standard deviation, or standard error unit. Since, in this specific example, the sample mean is 75, and the standard error of the mean 10, this means that there is a 68 percent probability that the population mean falls somewhere between 65 and 85. Further extension to two standard error units in each direction (±20) indicates a 96 percent probability that the sample mean differs no more than 20 points in either direction, and thus that the population mean almost certainly falls somewhere between 55 and 95.

In practical terms we are now in a good position to offer concrete estimates to the politician. We can tell him the risk he will run if he assumes that the sample is representative of the population. Provided he feels such a risk is warranted, he may begin his metamorphosis to Lawrence Welkian form.

The important feature of this example, however, is not that the politician takes a risk; it is that a precise statement of the amount of risk he must take is obtained *prior* to taking the drastic steps of plastic surgery and accordian lessons. Thus, knowledge gained from the sample is like a beacon shining into the murky future, making subsequent action more than a stab in the dark. This is the vitality of statistical inference.

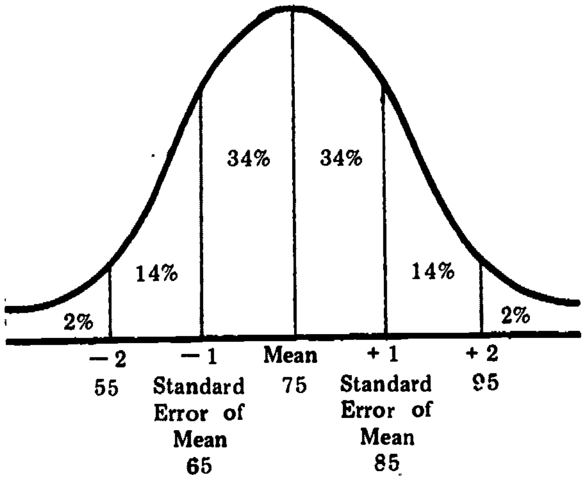
Earlier the standard error was signified by the formulation $\frac{\sigma}{\sqrt{n}}$ where n is the sample size and σ the population standard deviation. The value of σ, you will recall, is derived from the formula

$$\sqrt{\frac{\Sigma(M-x_i)^2}{N}},$$

where the symbol *M* represents the population mean. It is reasonable to assume that if the standard deviation could be supplied, surely the population mean (M) could also be supplied. The availability of the population mean completely destroys the

FIGURE 3

Normal Curve of Sample Means



impressive magic of the previous demonstration. It seems pointless to estimate the amount a sample mean will differ from the population mean when both sample mean and population mean are available and can be directly compared.

Naturally, the author is prepared for this criticism with still another miracle up his sleeve. It can be shown mathematically that as the sample size becomes somewhat large, say 40 or more, the population standard deviation (σ) is approximated by the standard deviation of the sample (s). Therefore, in cases where n = 40 or more, it follows that the standard error of the mean $\frac{\sigma}{\sqrt{n}}$ is closely approximated by $\frac{s}{\sqrt{n}}$.

And now the analysis is complete. After satisfying the conditions for the selection of a random sample, all further estimates are based on the sample. These procedures allow the researcher to predict what he will find in unknown worlds and at future times with a specifiable degree of precision, thereby eliminating the necessity for arduous, costly, and sometimes impossible exploration. The vehicle for these magnificent feats is, of course, the star ship "statistical inference".

FOOTNOTES

[1] Yerkes, R. M. "The Instincts, Habits, and Reaction of the Frog." *Psych. Rev. Monographs*, 1901, supp. pp. 17, 637.

REFERENCES

Anastasi, A. *Psychological Testing.* New York: Macmillan, 1968. pp. 39-70.

Games, P. A. & Klau, G. R. *Elementary Statistics.* New York: McGraw-Hill, 1967. pp. 249-250, 24-60, 168-209, 211-256.

McNemar, Q. *Psychological Statistics.* New York: Wiley, 1969. 1-4, 5-12, 31-40, 80-85, 41-54, 89-96.

Skinner, B. F. *Contingencies of Reinforcement.* New York: Appleton-Century-Crofts, 1969. pp. 144, 239.

Young, R. K. & Veldman, D. J. *Introductory Statistics for the Behavioral Sciences.* New York: Holt, Rinehart, & Winston, 1965. pp. 2, 6-7, 109-116, 9-16, 129, 137, 167-168, 172-174.

Yerkes, R. M. "The Instincts, Habits, and Reaction of the Frog." *Psych. Rev. Monographs*, 1901, supp. pp. 17, 579-638.

CHAPTER 8

Validity

TITCHENER'S BRAIN

The world abounds with lonely people. Often the college professor is among them. His relentless pursuit of the obscure and the esoteric frequently appears to be a preoccupation with tedious inconsequentials. It is not surprising that weeks of seclusion in his study or laboratory lead him to resort to almost any means which produce even brief moments of conversation or repartee. Doubtlessly, it was such loneliness that furnished the impetus for the inception of one of the most diabolical practices in social history — the "conversation piece".

A conversation piece is an object, curio, or relic which, once placed on the academician's desk or laboratory bench, evokes that first unfortunate question from an unsuspecting janitor, fire inspector, or misrouted sky diver.

Some pets make excellent conversation pieces. A hippopotamus or boa constrictor placed on the desk is almost certain to elicit a response from a passer-by. The discovery that carniverous plants will devour liverwurst sandwiches imprisoned for weeks in a desk drawer, has made them a favorite among those who insist on functionality even in conversation pieces.

But perhaps the most provocative conversation piece of all time was displayed in the office of Professor K. M. Dallenbach of Cornell University. There on his desk top in a large jar of formaldehyde was the brain of Professor Dallenbach's former mentor — the man who was chiefly responsible for the beginning of scientific psychology in America — William Bradford Titchner.

Titchener's insistence on rigorous scientific practice persuaded the American scientific and academic communities that human experience could be subjected to the methods of empirical science which had proven so successful in other disciplines. His scientific predilections survived even beyond the grave. It was his wish that his brain should become the property of Cornell University after his death; and thus, Professor Dallenbach, who had been student, graduate assistant, and later colleague and friend to Titchener, was entrusted with his brain. Dallenbach displayed this grisly memento in a prominent place on his desk where all could see the massive tumor which had caused Titchener's death.

At this juncture, the thoughtful reader is undoubtedly at a loss to discover how Titchener's pickled, tumorous brain has even the slightest bearing on psychological testing. As a matter of fact, you will remember that, while alive, Professor Titchener had often ridiculed the study of individual differences, on which psychological measurement is based, as sheer folly. One could not, he maintained, fruitfully engage in an effort which sought to delineate differences among people — on the contrary, the only proper course for the science of psychology was to attempt to describe and catalogue those features which men possessed in common. The uniqueness of an individual was irrelevant to the scientific study of a species of organisms.

Having opposed the testing movement so vociferously when alive, how could Titchener's brain, slumbering in its formaldehyde environment, launch us into a meaningful exposition of the basic tenets of psychological testing?

For some time before Titchener's death, his physicians had suspected the presence of the brain tumor. Changes in his coordination and the emergence of severe and, for Professor Titchener, unprecedented headaches, were among the symptoms which led the physicians to suspect a brain tumor.

The occurrence of a brain tumor is unfortunately all too common. It is not the reason, however, that Professor Titchener's case was introduced into what was ostensibly an exposition on tests and measurement. The relatively uncommon aspect was that after Titchener's death an autopsy was performed which revealed the presence of the tumor.

Tracing the chain of events we find: 1) Certain peculiarities or *differences* were observed in Titchener's behavior by himself and others. These differences were discerned with reference to the manner in which he had behaved and felt previously. The observation of these differences led him to seek the counsel of a physician. 2) The physician also observed differences in Titchener's behavior and appearance. These differences were discerned with reference to the behavior and appearance of other "normally healthy individuals". 3) The physician conducted medical tests which resulted in performance differences between Titchener and normal individuals. 4) On the basis of the differences, the physician made the diagnosis of "brain tumor". 5) Upon Titchener's death, an autopsy revealed the presence of a brain tumor, whose location and extent were consistent with the medical diagnosis.

The autopsy and verification of the tumor was a critical step in the procedure. Without it, all observations, test results, and diagnoses may have been totally irrelevant and inconsequential.

TESTS AND THE FUTURE

If you will now recite to yourselves the definition, previously presented, which defines a psychological test, you should find it relevant to this discussion. For the sake of the reader who, having mistaken this book for a girlie magazine, finds himself a disoriented, but nonetheless captivated, newcomer, we will reprint the definition below:

A SET OF STANDARD STIMULI AND PROCEDURES, INCLUDING ENVIRONMENT, INSTRUCTIONS, PRESENTATION OF TASKS, AND SCORING CRITERIA, WHICH, WHEN APPLIED TO AN INDIVIDUAL, YIELD STATEMENTS WHICH CAN BE USEFULLY RELATED TO HIS FUTURE BEHAVIOR.

Once again particular attention must be given to the final clause of the definition, ". . . which, when applied to an individual, yield statements which can be usefully related to his future behavior." All tests, whether they are performed by a physician, psychologist, or engineer are executed with the future in mind. The statements or test results are related to the future in such a manner that they will allow predictions to be made. These predictions, in turn, may indicate a necessity for certain actions or steps to be taken in order to prevent or, in some cases, insure a future outcome.

In the case of medical tests, results may indicate the active presence of a microbe or other pathological condition. Hopefully, results will thus lead to diagnosis which, in turn, will lead to medication or other treatment effective in eliminating the problem.

In somewhat uncommon instances, tests are constructed and administered to a population of subjects with no thought given to future action or events. Results of these tests may be placed in distributions or converted to statistics which do no more than describe the population. It was Francis Galton's dream that all citizens of the British Isles be given his psychological and anthropometric tests and these results be recorded for posterity. Such apparently pure scientific or academic endeavors always have implicit future implications, even if these be no more than to keep track of changes across time, or for comparison with other currently existing populations. But it is safe to say that the major purpose of a test (medical, psychological, metallurgic, aerodynamic, or whatever), is to predict future performance, conditions, or outcomes.

In some instances it may appear that a test is not so much predicting the future as it is predicting the current existence of a specific condition or situation at a location which is not observed directly. Such is the case when test results compel the physician to predict a growing tumor or microbe.

In all cases, tests predict specific conditions or outcomes. Tests are valuable if predictions made from them come to pass. To the extent that a test result is a good and accurate predictor, the test is said to be *valid*.

If you are new to the world of tests and testing, prepare yourself for repeated exposure to the term *validity*, or *test validity*. Validity is the whole ball game as far as tests are concerned. A test which has no validity is absolutely worthless, although it may persist as a relic, academic model, or monument to failure.

The advantage of the test is that of predicting from one time and/or space to another. Medical tests, such as those employed to detect malignant growths, allow for measurements to be taken at external and relatively accessible bodily locations. These measurements give us knowledge of existing physiological conditions in other, practically inaccessible, anatomical areas.

Early clinical psychologists and psychiatrists were associated with medicine and trained by physicians. It is natural that they would accept the "medical model" — which is that abnormal or deviant behavior is caused by some disease or physiological malfunction within the body. Early psychological tests developed by these clinicians were therefore constructed along the lines of medical tests: test performance was considered an external manifestation of internal pathological conditions.

More recent test constructors and developers do not assume any particular relationship between test performance and presumed internal, physiological, or even "mental", conditions. They see test results as measurements taken at one time and place which can be used to tell us how the individual is likely to perform at some future time and place.

It is apparent that whether taken as proof of an existing internal condition or simply as an index of later performance, the focus of the test is the future. It allows for predictions, thus giving its user the eyes which can see where his cannot yet see. But predictions can be incorrect, and the cost of acting on an incorrect prediction where drastic surgery, institutionalization, or great financial investment hang in the balance, can be dire. A test which does not allow the user to make accurate predictions can obviously cause only frustration, if not actual harm. Before a test can be used in any extensive manner where results are likely to influence or determine decisions entailing human resources, some prior knowledge of the probability that predictions will be accurate is an absolute necessity.

Procedures employed to determine the accuracy of predictions made from test results is called *validation*. For the time being, let us define *validity* as: *The degree or extent to which results of a particular test predict future outcomes or conditions in which the test user is specifically interested.* Procedures designed to assess validity are cleverly called *validation procedures.*

The post mortem examination is often an extremely critical step in the evolution of accurate medical tests and diagnoses. In Professor Titchener's case, diagnosis was validated when the autopsy revealed the tumor. You are perhaps relieved to find that the interest of the medical world in securing autopsy permission whenever possible is not merely ghoulish perversion, but a scientific necessity if better tests and diagnostic procedures are to be developed.

FACE VALIDITY

In some instances the very nature of a test may be such that all reason indicates that it must surely be related to the outcomes in which the test constructor is interested. When such a situation is evident, that is, when there appears to be an obvious relationship between test items and the performance or conditions the test user wishes to predict, the test is said to have *face validity*. This merely means that it looks like it should relate to the criteria of interest.

A bank vice-president of my acquaintance, for instance, told me about a "psychological test" which his bank uses routinely in selecting new administrative employees or bank officers. The prospective employee is taken out to lunch by those directly concerned with filling the new position, and when the food is placed before him, he is watched very carefully. If he tastes his food before salting it, this is considered an indication that he has good judgment, that he is likely to procede with caution and to be circumspect in all matters, including those involving money. If, however, he salts his food before tasting it, this is taken as evidence that he is impulsive and therefore would be imprudent or careless with bank funds. It is easy to see the rationale whereby my acquaintance developed this test. It seems a logical approach to the question and thus has "face validity". However, whether or not this procedure is effective in selecting good employees will probably never be answered. For if an employee fails the test, he is not hired and therefore no information as to how he might have performed will ever be available. If he passes the test, and is subsequently hired, it is probable he will work out to that reasonable satisfaction of his employers, since strict credentials and training are required before the prospect is interviewed.

Another example of good face validity, but questionable predictive efficacy, occurred in the selection of flight training candidates during World War II.[1] Since these candidates would eventually operate high speed aircraft, it seemed reasonable that their ability to operate the controls of simulated aircraft would have some relevance to success in pilot training and therefore to performance in real aircraft. As a result, most military aviation training facilities developed elaborate stationary machines which resembled the cockpit of an actual aircraft. Candidates were observed in terms of their general dexterity, reaction times, and the ease and speed with which they adapted to this situation. Not long after, research demonstrated that performance scores on the stationary airplane had little, if any, predictive value in selecting those individuals who would make the best pilots. It appeared that the factors which distinguished an individual who would become a good pilot had little relationship to the factors involved in operating the stationary aircraft. It was ultimately found that paper and pencil tests related to "personality", "intelligence", and other similar factors were far more effective in the selection of successful candidates.

Validity, therefore, is not a question of how things *should* turn out; it is rather an empirical statement of how they *do* turn out. Procedures must therefore be developed which will give information on the effectiveness of test results in predicting those outcomes for which the test is employed. Face validity may help a test constructor in selecting initial test items or questions, but eventually the predictive effectiveness of the test must be assessed in some straightforward manner. But if face validity is not enough — and it definitely is not enough — how is validity assessed?

For a moment let us drop back and recapitulate some of the more salient points which have been made previously. First of all, testing requires measurement. Usually the rationale whereby measurements are defined as one type or another depends upon the training of the individual administering the measurements, and thus, the purpose for which the measurements are made in the first place. One may observe a job foreman, physician, physical-fitness instructor, and psychologist each instructing their prospective employee, patient, student, or subject to lift a block of lead above his head ten times, increasing the weight in five pound increments on successive trials. All may measure the performance in exactly the same manner and record it in identical units, but the purpose of the test may be different in all cases. The job foreman's purpose is strictly to see if the prospective employee will be capable of performing the same task on the assembly line. He may call his test a "job capability test". The physician is concerned with the general health and strength of his patient and may call his measurements "muscle-tone measurements". The physical-fitness instructor refers to his measurements as "standing press". Finally, the psychologist may be interested in the coordination and concentration required to complete these tasks and may call his measurements "physical-concentration measurements".

Regardless of who makes the measurements, or for which purpose they were made, the crucial aspect to remember is that the measuring procedures *create differences.* Let us see how this notion of creating differences, on which we have been most insistent throughout the book, relates to the general concept of validity.

DEVIL WORSHIPPERS' DILEMMA

In order to make these concepts more tangible, a "real life" episode can best provide a ready vehicle. Let us assume that you are head recruiter for a local devil-worship cult. Your job is to find new candidates for Coven #402, an affiliate of the International Consort of Devil-Worshippers of America. Lately, a rash of first run movies released on late, late television have severely cut into attendance. Last week, for instance, when Birth of a Nation, starring George Washington and Jane Fra[illegible] was aired, even the young virgin who was to be the human sacrifice for that evening's Black Sabbath called to say she couldn't make it.

Of course, steps could be taken to remotivate the old membership, even such drastic steps as notifying the top man himself. But going all the way to the bottom seemed a bit drastic, and besides, His Majesty was hard to reach since he had received a new color TV from a patronizing Congressman. Rather than rejuvenate an old membership already pulpy soft from overexposure to Sunrise Sermon, it seemed more expedient to bring in new blood, so to speak. Thus, your task as head recruiter came into existence. All manner of recruiting possibilities are considered: full-page advertisements in the *Wall Street Journal,* a listing with the State Employment Service, billboards, leaflets dropped from helicopters, etc. Ultimately, all are rejected as being too indiscriminate and "devil-may-care".

No, this time careful selection procedures will insure that only the best candidates are encouraged to seek membership. Quite naturally, the possibility of constructing a psychological test to aid in this selection procedure comes under consideration. But what sorts of questions, puzzles, tasks, or problems should be employed as test items? By now two requirements of any group of items selected occur to the reader.

1. The test items, taken as a group, must be such that, by and large, different individuals achieve different test results or scores. This is merely a restatement of our basic notion that a test, as a form of measurement, creates differences. In previous pages it has been noted that the more differences a test creates — that is, the greater the extent that different individuals achieve different test scores — the more likely it is that the test will relate to future events or situations.

2. Test items should be selected which are likely to relate to the particular outcome or condition in which the test constructor is interested.

Cognizance of these two requirements, however, offers little help in selecting questions, puzzles, tasks, etc. which will be efficacious in screening good coven membership prospects from poor ones. At this point, the test constructor must rely on his own devices, and this generally means that he will select test items, at least initially, in terms of face validity. He will experiment with tasks or questions which look to him, by some rationale, or logic, like they should relate to the particular performance in which he is interested. to solve problems or behave in ways similar to current "good" or exemplary members would make effective test items.

But no matter how "obvious" the relevance of prospective test items to the future outcome, it must be established that these items actually can be employed to predict this outcome. As you, the reader, have already been apprised, the process by which this is established is called *validation.*

Proceeding on intuition, logic, or reason, the coven recruiter selects a group of problems and questions which appear, on the basis of face validity, to be extremely relevant to exemplary participation in Black Sabbaths, supplication, incantation, diabolic worship, and sacrifice. Those who would be best suited for membership should have no difficulty in answering the questions and in performing the tasks which have been selected as possible test items. They should achieve the highest test scores.

The group of items is compiled in the form of a tentative test and given to ten individuals who have recently applied for membership. Subsequently, the test is graded and a test score recorded for each individual. The results are presented in Table I.

TABLE I

Prospect Name	Test Score
Carla	92
Ricardo	216
Bruce	114
Milford	401
Prudence	32
Fred	56
George	155
Simone	170
Leona	172
Brian	5

A brief inspection of Table I is very encouraging to the coven recruiter. The first requirement we mentioned earlier has been satisfied — all individuals have achieved a different score. Furthermore, these differences are distributed somewhat evenly in a range from 5 to 401. As the reader already knows, such a distribution of outcomes at least has a chance of being related to the future outcomes the test constructor has in mind.

The very fact that a need for selecting only some of those who apply (with the purpose in mind of ending up with only the best members), implies that some criteria or standards already exist whereby members may be judged as desirable, less desirable, mediocre, or miserable. While it would seem to be a necessity that any club, organization, or business have a precise, standard, and regular way in which members or employees are evaluated, it is unfortunately true that in most cases such evaluations are haphazard and highly subjective, relying on impressions and opinions of supervisors or leaders.

The coven recruiter receives another break here. For some time, objective performance reports have been compiled at six-month intervals. These performance reports are expressed in a quantitative score from 0 to 100. These performance rating scores emanate from all aspects of devil-worship and are quite objective. Naturally, those performing in the manner most desirable to the leadership receive the highest score. After administering the preliminary form of the test to the ten new applicants for membership, results cannot yet be employed to accept some and reject others for official membership. This would assume that face validity was enough. We have already made it clear that this is not the case. If a test is worth its salt, it must predict. Not until its predictive power is validated should it be employed to the advantage or detriment of any individual. Such a procedure would only perpetuate an error which is historically and even currently all too prevalent.

For the practical purpose of validating the test, all ten applicants must be accepted for membership. At the end of six months, the usual performance ratings for all members, including the ten neophyte Satan children, are compiled. Now results of the tentative test can be compared directly to performance ratings taken at the six-month assessment. This comparison appears in Table II.

TABLE II

Member	Test Score	r = +.90	Six-Month Performance Score
Carla	92		37
Ricardo	216		86
Bruce	114		57
Milford	401		97
Prudence	32		20
Fred	56		39
George	155		75
Simone	170		77
Leona	172		98
Brian	5		10

A cursory inspection of the two sets of scores thrills the coven recruiter all the way to his black heart. It appears that there is a definite relationship between the test score and the performance score. Certainly the higher test scores are associated with the higher performance scores, indicating that those doing best on the test also tend to become the best members.

COEFFICIENT OF CORRELATION

Visual inspection of two sets of scores, even where as few as ten individuals are concerned, can be confusing. Trends which seem obvious are sometimes more apparent than real. Furthermore, it is one thing to discern a relationship, but if predictions are going to be made from such a relationship, it is necessary that the extent or degree of the relationship also be known.

Once again it is time for mathematics, and particularly statistics, to aid in both discerning and describing such relationships. A descriptive statistic particularly suited for relating two sets of observations is called the *coefficient of correlation*, or often just *correlation*. A correlation is just that, a *co-relation*, indicating that two sets of scores are related to each other.

A numerical statement of such a co-relationship is called an index, or coefficient, and results when the appropriate statistical calculations are applied to raw numbers or scores. There are many types of correlation coefficients. The high priests of numbers insist that some statistical treatments are indicated for scores or outcomes of one type and other statistics for different types of scores.

The type of correlational procedure most widely employed in psychological testing was formulated by the British mathematician Karl Pearson and became known as the Pearson Product Moment Coefficient of Correlation. Like others, the Pearson Correlation results in an index or coefficient which may range from +1.00 to −1.00. Thus the coefficient is really a statement of the relationship between two sets of observations. Suppose that a detective finds a blond hair at the scene of the crime. Provided that the hair color is not due to a hairdresser's artistry, what color of eyes is the suspect most likely to have? You may guess blue — and that is probably a wise guess. But why guess when concrete data are available? We could simply record the eye color of a large sample of naturally blond subjects. A correlation between the two observations (hair color and eye color) can then be performed and will result in a coefficient of correlation which could assume any value between +1.00 and −1.00. The plus or minus indicates the direction of the relationship; the magnitude of the coefficient indicates the degree of the relationship. If every time we observe a blond we also observe blue eyes, we will get a correlation coefficient of +1.00. This is called a perfect positive relationship. If correlational procedures had yielded a negative perfect correlation, a −1.00, this would have meant that once a blond was observed we would have known for certain that the eye color would not be blue. Once convinced that the criminal was blond all blue-eyed people would immediately be beyond suspicion.

The magnitude of the correlation can range from −1.00 to +1.00. A coefficient of .00 indicates no relationship at all. If observations of hair and eye color resulted in such a coefficient of correlation, this would mean that knowledge of hair color is of no help in predicting eye color. Although the coefficient of correlation can range between .00 (no relationship) and ±1.00 (perfect relationship), either of these extremes is rarely observed. Most correlations fall somewhere in between these two values. Another feature of the correlation which should be noted is that it is *reciprocal*. One can as easily predict hair color given eye color as one can predict eye color given hair color.

But to return to the plight of the recruiter for Coven #402. You will recall that he has in his possession two separate sets of scores of his ten new candidates. One set resulted from the tentative test which hopefully will become a valuable predicter of acceptable new members. The other set of scores came from performance ratings made during the first six months of participation in coven activities. Table II depicts these two sets of scores along with a correlation which has been computed between them. In this instance the Pearson Product Moment Coefficient of Correlation seemed an appropriate correlational statistic to employ. The result of this statistical analysis from Table II is r = +.90 (*r* is the coefficient of correlation).

The fact that the correlation is above .00 in magnitude tells us there is a relationship between test scores and performance scores. The fact that the coefficient is preceeded by a plus sign (+), tells us that the relation is positive, and therefore, that those doing well on the test will also tend to receive a good performance rating.

The *extent* to which the test score is related to performance ratings is important. If the correlation is high, approaching +1.00, the predictions about positive future performance ratings are likely to be accurate. Applicants for membership can be accepted or rejected according to their test scores. Naturally, such a practice will not only allow for the best prospects to be accepted, but will also eliminate the cost of weeding out individuals who have proven themselves unsuitable members.

As already noted, a perfect correlation is rarely observed, and we have to deal with a more or less acceptable magnitude. The acceptability of the magnitude is determined largely by practical considerations having to do with the purpose for which the test is employed and the relative cost of making a wrong prediction.

Taking the current instance as illustrative, Table II indicates a correlation of +.90. Let us demonstrate how a test which correlates with performance at this magnitude can improve over current membership selection practices. We find from past records that approximately 50% of all new members were judged not suitable because they did not get a score of 70 or above on their performance rating after the first 6 months of coven membership. The cost of recruiting, training, and maintaining a new member up to the six-month point is $5,000 per individual. In the past five years alone 2,000 new members have been so processed, and 50% of these did not make the acceptable 70 performance score. Since each member costs $5,000 to process, it follows that during this period

$5,000,000 (5,000 x 1,000) has been lost due to the ineffective selection procedures.

Let us now examine how the test with its correlation of +.90 will improve on this. Recruiting practices in the past operated on "the more the merrier" policy; perpetually requiring large numbers of new recruits to replace those who were "fired". By employing the test as a selection device, only those who obtained the highest scores would be accepted. But the coven recruiter must know how high the score should be to consider an individual for acceptance and what percentage of these individuals will ultimately prove to be acceptable members. It would be most economical if all who were accepted for membership became regular, card-carrying, dependable members. Thus training monies would never be wasted on individuals who later were dropped because of inadequate performance.

The information in Table III is exactly what the recruiter needs. It is relevant to the current situation where in previous years all applicants were taken into the training program, but only 50% proved acceptable.

Proportions of applicants accepted appear across the top row and range from .05 to .95. An acceptance criterion of .05 means that only those scoring in the top 5% will be accepted. An index of .10 means only the top 10% of those individuals applying will be accepted, and so on to the .95 proportion.

The left hand column of Table III presents coefficients of correlation (r), ranging from .00 to 1.00. Moving down this column to .90, we find the magnitude of correlation which the coven recruiter obtained between his preliminary test and six-month performance scores. Moving across the row at this point to the .05 column we will note that if only those applicants scoring in the top 5% are selected, 1.00, or 100% will prove to be adequate. Using the test as a device for screening out all but 5% of the applicants should result in a group of new recruits who all will be excellent coven members. Moving further across the row to .10, the data indicate that relaxing the acceptance to the highest 10% will still result in perfect selection. When 20% are accepted, only 1% will not work out, etc. Thus Table III demonstrates some relationships which must be considered in terms of the practical requirements of any recruiting or selection situation.

It is apparent that when a test which correlates at a given magnitude with successful performance is employed as a selection device, the percentage of highest scorers which should be accepted depends on purely practical consideration. Such things as the cost of accepting candidates who do not succeed, the abundance of applicants, the urgency for new members or employees, etc. determine the final decision about the magnitude of the score required of a new member.

ACCEPTABLE CORRELATION MAGNITUDE

At the risk of repetition and redundancy, let it be said that the magnitude of a correlation is a practical issue. Table III demonstrates the validation process which shows that for any percentage of applicants accepted, an increasingly greater number can be expected to succeed as the coefficient of correlation increases.

In a realistic manner, it is possible for a test constructor to decide in a very direct manner whether or not the test he has engineered will be of any practical value to him. The magnitude of the correlation between the test and the performance criteria must be such that the predictive force of the test will outweigh the bother of test administration and scoring, and will also offset the realistic cost of error of prediction.

Table III reveals that if the top half of the applicants (in terms of test scores) are accepted, a test which correlates only .50 will result in 67% success. This is an improvement of 17%. If an improvement of 17% in employee or member selection

TABLE III

Selection Ratio
Proportion of Employees Considered Satisfactory = .50

r	*.05*	*.10*	*.20*	*.30*	*0*	*.50*	*.60*	*.70*	*.80*	*.90*	*.95*
.00	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.05	.54	.54	.53	.52	.52	.52	.51	.51	.51	.50	.50
.10	.58	.57	.56	.55	.54	.53	.53	.52	.51	.51	.50
.15	.63	.61	.58	.57	.56	.55	.54	.53	.52	.51	.51
.20	.67	.64	.61	.59	.58	.56	.55	.54	.53	.52	.51
.25	.70	.67	.64	.62	.60	.58	.56	.55	.54	.52	.51
.30	.74	.71	.67	.64	.62	.60	.58	.56	.54	.52	.51
.35	.78	.74	.70	.66	.64	.61	.59	.57	.55	.53	.51
.40	.82	.78	.73	.69	.66	.63	.61	.58	.56	.53	.52
.45	.85	.81	.75	.71	.68	.65	.62	.59	.56	.53	.52
.50	.88	.84	.78	.74	.70	.67	.63	.60	.57	.54	.52
.55	.91	.87	.81	.76	.72	.69	.65	.61	.58	.54	.52
.60	.94	.90	.84	.79	.75	.70	.66	.62	.59	.54	.52
.65	.96	.92	.87	.82	.77	.73	.68	.64	.59	.55	.52
.70	.98	.95	.90	.85	.80	.75	.70	.65	.60	.55	.53
.75	.99	.97	.92	.87	.82	.77	.72	.66	.61	.55	.53
.80	1.00	.99	.95	.90	.85	.80	.73	.67	.61	.55	.53
.85	1.00	.99	.97	.94	.88	.82	.76	.69	.62	.55	.53
.90	1.00	1.00	.99	.97	.92	.86	.78	.70	.62	.56	.53
.95	1.00	1.00	1.00	.99	.96	.90	.81	.71	.63	.56	.53
1.00	1.00	1.00	1.00	1.00	1.00	1.00	.83	.71	.63	.56	.53

Source: Taylor and Russell (32).

results in appreciable savings ot resources, the test will be valuable.

On the other hand, a savings of 17% may not be worth the time, effort, and cost incurred in test administration alone. However, by consulting a table such as Table III, a test constructor can determine exactly what various magnitudes of correlation offer at a given acceptance percentage in terms of savings over current selection procedures. If the preliminary test he has constructed does not correlate highly enough with the performance criterion, it may well be a case of "back to the drawing board". Changing items, inserting more or different test questions or problems may result in a test which allows for increased validity and therefore may make its use worthwhile.

Often tests which correlate extremely poorly with performance criteria are immediately dismissed as worthless. A test which correlates .10 does not appear impressive. Such a test would increase selection accuracy from 50% to 51% in a situation where 90% of all applicants were selected and only 10% rejected. But a savings of even 1.0% could amount to thousands, or even millions, of dollars in resources in some real life situations. Instances where large numbers of individuals are and must be regularly selected, and where inaccurate selection is costly, are situations where any improvement in selection, however slight, is likely to be of tremendous practical value. Certainly, the one-eyed man is king in the valley of the blind, and likewise, a test with slightly improved validity is a valuable asset.

VALIDATING CRITERIA

In its simplest form, the validation of a psychological test involves relating results ensuing from the test to other observations or scores which are presumably inalienably associated with the event, condition or situation of which accurate prediction is desired. The observations, scores, or performance data which characterize the situation we wish to predict are called *criteria*. A single set of performance data may thus be referred to as a *validating criterion*, and multiple types of observations which serve the same purpose are, of course, *validating criteria*.

In a previous example the coven recruiter employed a validating criterion which was standard and highly objective. Would that the real world were so sublime! Undoubtedly the single greatest difficulty in constructing a psychological test is finding a suitable criterion or set of criteria against which the test can be validated.

Since, eventually, some type of statistical procedure — such as a correlation — will be employed to identify and express the degree of relationship between test results and criterion, a criterion which is numerical in form is desirable. But how many employers or businessmen of your acquaintance will reply with a numerical answer when asked how well a particular employee performs his job, or otherwise contributes to the overall success of the enterprise? Generally your question will either provoke a nebulous statement such as "pretty good" or "fair", or it will prompt a two-hour fillibuster brimming with anecdotes or specific instances recounting the past behavior of the employee in question. Neither of these responses will be of any practical value to you as a test constructor. You are looking for a highly succinct but relevant statement. How sweet a "10", "87", or "232" would be!

But a number, like any verbal statement, must bear a particular descriptive relationship to behavioral occurrences or performances if it is to be brought into alignment with other sets of observations or test results. Therefore, not just any old set of numerical statements which are handy can be fruitfully employed as a validating criterion.

There are many kinds of physiological measurements which yield numerical statements. Blood count, heart rate, and vital capacity are just a few. A researcher impressed with the specificity of these measurements may be tempted to employ them as criteria against which his pregnancy test can be validated. Unfortunately, regardless of how objective and numerical these measures are, a physician will tell you that they are in no way associated with pregnancy. The researcher must search for other measures.

WHO IS PREGNANT?

The allusion above to pregnancy gives us an opportunity to make several additional points about criterion selection. In the end a psychological test, or any test for that matter, is a practical invention. The same purpose or motivation which provided the impetus for an attempt at constructing a test should also provide the selection of a validating criterion.

Many individuals are interested in the fact that a woman is pregnant. Strangely enough, often those most keenly interested have never so much as nodded to, or shaken hands with, the woman in question. In the past, pregnancies have led to births, and births are of practical importance to hosts of individuals including the prospective father, baby crib manufacturer, university president, and yes, even the local funeral director.

Of course, others are interested in even those pregnancies which do not result in births. The physiological changes occurring at any point in pregnancy are extensive and can threaten health. The physician is interested in all pregnancies. The social and psychological implications of pregnancy are extensive in our culture. Teachers, clergymen, sociologists, economists, and parents of teen-age children are likewise tuned in to pregnancy.

The fact that many persons exhibit an interest in a particular phenomenon such as pregnancy, does not indicate that they do so for the same reason. While all individuals would receive valuable information if they knew exactly what woman was pregnant, the specific situations under which the condition was acquired, the time of conception, etc., would not be judged relevant by those professing a general interest. The sultan may wish to know only who in his harem is pregnant at a given time.

When considering a criterion against which a test will be validated, the very specific purpose or need which led to the construction of a test should be the guiding principle.

A manufacturer who wishes a test to aid in the selection of employees must often become aware of precisely what his interests are before he can begin to provide the test constructor with an adequate criterion. If the manufacturer's concern is production, a test which selects individuals who work in such a manner that they produce the most, appears reasonable. A possible validating criterion would be the number of parts produced per hour or day.

An employee's actual production is not unrelated to a myriad of other factors, however. He must first appear at the factory regularly and on time. Since he works with others his ability to cooperate, and in general, realize the rights and prerogatives of his co-workers cannot be disregarded. Poor adherence to safety standards, abuse of expensive equipment, and dishonesty are all performance outcomes that are not conceptually related to production; but, nonetheless, can bring it to a complete standstill, or otherwise engender conditions which will destroy the business or industry.

It is rare that the criterion to which test results will be compared can be as unilateral as production. Most validation procedures employ multiple criteria, or single out a criterion which is really an amalgamation of multiple factors.

Perhaps it would be helpful at this point to outline several desirable features of validating criteria.

1. As mentioned above, the criterion or criteria must be inalienable from the purpose for which the test was constructed. A test which is constructed for the purpose of detecting brain tumors must ultimately be validated by comparing test results to actual incidence of brain tumors. A test which selects "better" employees

must be validated against all criteria which will accurately define a "good" employee.

2. Validating criteria must be objective in the sense that there is no difficulty in discriminating between those who succeed and those who fail. A patient either has a brain tumor or he does not. An employee performs satisfactorily or he is unacceptable. Unless the validating criterion is unambiguous, specific, and precise, it is of little use in test validation.

3. Validation may take place if the validating criterion occurs in a discreet (two-choice) language such as pass-fail; acceptable-unacceptable; tumor-no tumor; disease-no disease. However, the statistical procedures by means of which correlations are computed between test scores and criterion data must be other than the Pearson-Product Moment Procedure we discussed earlier.

All things being equal, validating criteria which allow for the occurrence of a wider range of differences in performance will allow for greater validity (higher correlations) to be evidenced. Statements occurring in the validating criterion are measurements. Discussion in previous chapters has made it clear that measurement procedures which allow for the greatest number of different and specific outcome statements are superior. Of course, measurements which yield number statements of a continuous and wide range are likely to be most advantageous, and convenient.

4. Once a test is validated against a specific validating criterion, the resulting correlation, and thus validity, cannot be extended to other situations unless further considerations take place.

As an example, a test which correlates quite highly with college success at a particular university may prove of great utility in the selection of students from among those applying. The test results may not correlate well at all with college success at another institution, although the curriculums appear similar. Thus validity cannot be too quickly generalized from specific instances regardless of the magnitude of the correlation obtained. In the final analysis, validation must occur in each and every new instance a test is applied. Test validation is an ongoing, ever-extending process. This ongoing, or growing, feature of validation is often called "construct validation".

TYPES OF VALIDITY

There are numerous sources and types of performance data, or observations, which could be employed to validate a selection of best items. Those finally utilized should be so chosen with the four points noted in the previous section well in mind. Further determinants of the final selection of validation criteria will depend on practical consideration such as convenience and experience.

Of the many validating criteria and validation procedures possible, several broad groupings, or types, of validity are prevalent, and therefore, traditionally noted in books dealing with psychological tests and measurements.

CONTENT VALIDITY

The content validation is probably the least complicated type of procedure. It may be summed up by the statement: If you want to know if a man can do the job, ask him to do it!

In a content-validation situation, test items or problems constitute the same type of task or performance comprising the validating criterion.

A typing test will most likely measure an array of varied materials to be typed within a strictly enforced time limit. If the validity of the typing test is established by correlating test results with subsequent, on-the-job, typing output, this is an example of content validation. In this particular instance both test results and performance data would undoubtedly consist of speed and accuracy of typing.

Although the general activity required by the test does not differ substantially from that observed in the validating performance, the test typically is structured so that the full gamut of materials or problems which are likely to be encountered on the job are represented. Information gained from a brief, well-structured test can save much effort and expense when employed as a selection device.

Content validation is the best approach to validation when individuals already possess the specific skill or ability which will be required in the future situation, but it is a question of degree, or extent, of attainment. All but a few states now require a driving test in which the applicant is asked to drive an automobile under controlled and structured conditions. Presumably such tests are content-validated by later driving records. More often drivers' tests proceed on the force of their strong face-validity.

Performance situations or jobs which involve a highly complex and specialized behavioral repertoire must often provide a long training period for new employees before they can adequately behave in the manner required. Such situations are not appropriate for the construction of a test which will be content-validated.

Applicants for a position as an astronaut could be subjected to a job selection test which involved piloting a space ship to the moon. The test would be validated by correlating the degree of success on the test mission with performance on subsequent missions. Since even minor errors could result in the loss of both applicant and space ship, an attempt at content validation would be most expensive. In addition there could be no validating criterion if all applicants did not outlive the completion of the test.

Content validation often succeeds because of the high degree of face validity between test and validating criteria. A typing test seems a "natural" in the selection of girls for the typing-pool. However, if one succeeds in the typing-pool for other than typing proficiency, scores on the typing test may correlate poorly with job success. Regardless of face validity, a typing test may not help in the selection of job applicants. Other tests and other validation approaches may be more appropriate.

CONCURRENT VALIDITY

Tests may be validated with the use of existing information ensuing from other sources. Information against which the test is validated should be up to date having been compiled either slightly before or after test administration. Thus the term "concurrent" is applied to this type of validation.

Grades, scores on achievement tests, performance ratings, and outcomes on qualifying or comprehensive examinations have been employed as concurrent validating criteria.

One prevalent criterion employed in the validation of psychological tests is the score from a pre-existing test which proports to measure the same thing. Scores from a new test constructed to assess knowledge of American history may be correlated with scores from an older, previously validated test.

Constructors of new intelligence tests almost invariably validate their test with scores from either the Stanford-Binet Intelligence Test or the Wechsler adult or child intelligence scales. In some instances, both Binet and Wechsler scales are employed as validating criteria.

This practice of employing an existing test as a standard against which a new test is compared and validated often confuses the student. The impetus for constructing a new intelligence test is in the American tradition an attempt to construct a "better" one. But how can a new test be "better" if it is compared, and, in fact, validated against those which it wants to top? Surely, one cannot do better than that which is held out as the ultimate criterion.

The resolution of this apparent contradiction is that there are numerous aspects in which one thing may be said to be better than another. The time required to administer the Binet or Wechsler scales may exceed an hour, and must be given on an individual basis. A new test requiring only five minutes and administered to millions of individuals, simultaneously, over television, is obviously "better" in a practical sense than the older tests. Inasmuch as scores on the new test correlate highly with scores on the old tests, the tests constitute equivalent measurements, and they are therefore equally valid as intelligence tests.

CONSTRUCT VALIDITY

In psychological testing, as in all other areas of human discourse, one must constantly separate what people do from what they say they do. Construct validation of a test is a case in point.

Scientifically speaking, a construct is a term which is used as a convenience to describe a large group of individual occurrences or phenomena which share common characteristics. Gravity is, in this sense, a construct. Gravity is a description of what has happened in the past when wide varieties of objects have been released without support in a wide variety of environmental locations and situations.

As a matter of convenience, however, the term gravity is offered to predict what will happen in the future, and, at the same time, identify this future occurrence as belonging to a class of similar occurrences which have happened in the past. The book will move to make contact with the floor when it is pushed from the desk, you predict. You may further say it will do so because of gravity.

The reason that the term gravity is useful to us is that it allows us immediately to relate large groups of occurrences, both past and future, and actually refer to them with one short, simple word. "Gravity" saves wear and tear on our vocal cords.

But gravity is not something we will touch, feel, bite, chew, or see. It does not exist in the same sense as an apple or your pet goldfish, nor does it emanate from some specific geographical location where a giant gravity generator has been cleverly disguised and hidden.

When we see things repeatedly fall toward the earth, we assume that some force has motivated them. The term gravity is not as much the name of this force as it is a description of the fact that things fall toward the earth.

When different people are observed to behave with consistent similarities in similar situations, and when we see no strings or other external inducements pulling them, in the past we have concluded that some common force inside them caused this behavior to occur. Obviously, if all people behave this way, all have the force within them. If different individuals do it to a different degree, then they have a different intensity or extent of the force. If some do it not at all, they do not have the force, or it is dormant, or too weak to be effective.

From observations that individuals appear to do the same things in what appear to be similar situations, psychologists have historically assumed the operation of a common force within them. Like gravity, this force is never seen or felt directly and as such is truly a construct. Some test constructors design tests which are supposed to measure or detect the presence of some force or construct. Such tests are validated by a procedure called *construct validation.*

The following steps outline construct validation.

1. The test constructor observes similarities in the behavior of individuals which are consistent from person to person. He assumes that this behavior pattern is due to the operation of a force which is differentially present or at least, differentially functional in different people.

2. Also, from his informal observation the test constructor makes some guesses as to what the force (which he cannot see) must be like; its intensity limits (which he cannot measure); and its peculiarities. These hunches are used to select test items and problems. They are chosen because they are believed to reflect the presence and extent of the construct as it operates in various individuals.

3. Hypotheses, or guesses, as to how different groups or populations of individuals will score on the test are formulated in terms of the test constructor's assumptions about the nature of the force or construct.

4. Hypotheses are tested by giving the test to these selected populations or groups. If the hypotheses prove to be correct, the construct (the test constructor's assumptions about the nature of the force) are confirmed and the construct is to this extent "validated". Subsequent hypotheses are put forward and similarly tested. If confirmed, the test is further validated.

5. If hypotheses are not confirmed, assumptions about the nature of the force, or construct, are altered to fit these outcomes. The construct is thus changed; items may be added to, or deleted from, the test; and new hypotheses generated and tested.

6. Construct validation is never completed. Beliefs about the nature of the construct are constantly changed, or solidified, as hypotheses are confirmed or rejected. Test items are accordingly changed.

A test publisher will typically publish all of the various research projects which have, to date, transpired in construct validation of his test, along with the rationale for the generation and testing of each hypothesis.

It is difficult to criticize the practical outcome of a construct validation procedure. A test is administered to a variety of different populations in a variety of environmental settings. This broad exposure, and the resulting outcomes, provide a potential test consumer with ample data to make an intelligent judgment on the feasibility of using the test for his particular purpose. In a real sense, the more specific instances in which results are correlated with validating criteria of diverse types, the more we know about a test, and that to which it relates. From this standpoint, we know more about the test's validity.

Initially, it was stated that one must separate what is done from that which is said about what is done. In the past, those who have employed construct validation have obviously been more concerned with proving the existence of the construct than they have been with providing a test which will allow for predictions, which have any practical advantage, to be made.

Regardless of how a new psychological construct is advertised, remember that it exists only inasmuch as independent observations or occurrences reveal common relationships. As these individual occurrences become larger, and their specific commonalities specified, the construct becomes more formidable and useful.

A good point to remember when considering psychological constructs is that we did not go searching for gravity as much as it came to greet us. Gravity became a construct because eventually it was convenient to label the myriad of similar occurrences with a simple summary term. Constructs which have proved valuable in other scientific situations have developed in the same manner — primarily, that of expediency. The great gold rush to discover the construct appears unique to psychology and psychological testing.

FORMAL PRESENTATION OF VALIDITY

When psychological tests are constructed, their use may be restricted to a particular business, educational setting, or industrial concern. The individual who developed the test may have done so with a specific purpose in mind which is of interest only to him. This does not exclude the possibility that the test will be of value to others. Such value must not, however, be assumed. An employer or personnel director would be foolish to continue screening applicants with a testing device which had not been proven valid in his particular business or enterprise.

But the construction of tests is often part of an independent industry which developes, validates, publishes, and sells tests to educational and industrial concerns. Tests generated with the notion of wide marketability must be cognizant of engineering tests to have widely established validity. Often these tests will be validated against a wide range and number of validating criteria, and will concurrently be validated against pre-existing tests which apparently attempt to provide a similar predictive or selective function.

When presented to the public, which is expected to buy the test, the publisher supplies complete information on how the test was constructed, including the manner in which it was validated and the resulting correlations. Often these published correlations are referred to as the test's "validity". Later, users of the test may carelessly say, for example, the Jones Reading Readiness Test has a validity of +.87.

Although a correlation of +.87 is a relatively high one as correlations go, such a statement, taken in isolation, is meaningless when the practical decision of whether or not the test should be employed in a particular setting is considered. Validity is a specific determination. A correlation of .97 or even 1.00 does nothing to recommend the use of the test unless the validating criteria which were employed to generate the correlations are comparable to the situation in question. The test manufacturer has done all that he can when he publishes, in specific and precise detail, the validating procedures employed by him and others who used the test along with resultant correlations. This information should give the consumer some indication as to the feasibility of "trying out" the test for his purposes. No matter how comparable his situation seems to that reported by the test publisher, he must ultimately validate the test in his setting.

VALIDITY — ENOUGH SAID

We have rambled far and wide from Titchener's brain in a jar, always attempting to elucidate the concept of validity, and, specifically, how it pertains to psychological testing. There is more we could say, and indeed in the next chapters, validity will not go unmentioned.

Before we can go further we must introduce another concept, as inalienable from measurement and testing as the Pope is from his religion. This is the concept of *reliability*. Traditionally, reliability is dealt with before validity. Before you judge the current change of order as just a further demonstration of the author's perverse nature, read on where undoubtedly his wisdom and forethought will again prevail!

FOOTNOTE

[1] Melton, Arthur W. (editor) *Apparatus Test*. Washington: Government Printing Office, 1947.

REFERENCES

Anastasi, A. *Psychological Statistics*. New York: Macmillan, 1968, 3-8, 99-157, 72-76, 21-30.

Cronbach, L. *Essentials of Psychological Testing*. New York: Harper & Row, 1970, 22-26, 115-150, 128-137.

Freeman, F. *Theory and Practice of Psychological Testing*. New York: Holt, Rinehart & Winston, 1962, 5-7, 88-117, 34-38.

McNemar, Q. *Psychological Statistics*. New York: Wiley, 1969, 125.

Page, J. *Psychopathology*. Chicago: Aldine-Atherton, 1971, 5-11.

Taylor, H. & Russell, J. "The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection: Discussion and Tables. *Journal of Applied Psychology*, 1939. 23, 565-578.

Ullman, L. P. & Krasner, L. *Case Studies in Behavior Modification*. New York: Holt, Rinehart & Winston, 1965, 2-26.

CHAPTER 9

Reliability

THE BOBSEY TWINS — VALIDITY AND RELIABILITY

In the last chapter the concept of validity was introduced along with logical and statistical procedures whereby it is typically assessed. It was mentioned that validity was a matter of constant and continuing concern in the construction and application of psychological tests. Another term which is equally inextricable from any consideration of psychological testing is *reliability*. In fact, where measurement of any type is concerned the Bobsey Twins, validity and reliability, are foremost on the scene. Let us learn more of this deadly duo.

To begin with, a summary highlighting some definitions and relationships appearing in previous pages is appropriate.

1. Measurement was defined as a set of devices and procedures which create differences when applied to particular individuals or aspects of the environment. Whether a psychological test, ruler, thermometer, or micrometer, the measuring device, along with the procedure by which it is applied, creates a population of results, scores, or outcome statements which differ from each other.

2. Measurements may or may not be of worth. If scores, readings, or test results are to aid us, they must relate to other scores, measurements, or observations which are relevant to our purpose. Ultimately, they must help us predict some aspect of the future.

3. If a set of measurements such as test scores are related to other events, the occasion of which we are interested in predicting, then presumably the administration of the test, and the ensuing results will allow us to predict. A test whose results are related to other events in which we are interested is said to be *valid*.

4. A relationship between a set of test scores and some criterion is, typically, both discerned and described by a statistical procedure which yields a statement of the relationship called a *coefficient of correlation*, which specifies both magnitude and direction of the relationship.

VALIDITY AND SINGLE INSTANCE

Presume that a sales director has generated a population of test items which he hopes can be related to performance as a salesman. Naturally, his aim is to discover a test which can be employed to select those individuals most likely to succeed.

The tentative test is administered to a group of 10 individuals who answer a newspaper advertisement describing the characteristics of the job. After taking the test all 10 are hired, and begin work the following week. Their performance on the job is evaluated daily, the primary observations being the number of sales produced. Thus, at the end of 30 days, sales data are compiled and summarized for each of the 10 new employees. These sales data and the score achieved on the tentative test are presented in Table I for each of the 10 individuals.

TABLE I

Applicant	Test Score	Total Sales For First 30 Days
Jones	125	630
Smith	80	400
Brown	22	110
White	47	235
Peters	216	1080
Curry	98	480
French	50	250
Black	72	360
Lawrence	31	151
Tyler	65	325

r = +1.00

A Pearson Product Moment Coefficient

is computed between test scores and sales, and is found to be a fantastic +1.00 — indicating a perfect positive relationship. Those scoring highest on the test have made the most sales. Quite clearly, if test scores had been employed as a selection device, those scoring low on the test could have been eliminated from further consideration with considerable savings of time, money, and embarrassment.

From this single administration of the test to a group of 10 applicants and a correlation of test results with sales output, it appears that the sales director is on the right track. Now he can use the test to select all future employees and save his company a bundle of money. He can send the test to other branch offices throughout the country and probably become the youngest vice-president in company history.

TEST, RE-TEST DISCREPANCY IN RANKING

But the sales director, wisened by years of watching television commercials, is uneasy. It seems too quick and simple. To satisfy his curosity, he does a peculiar thing. He re-administers the test to the 10 new employees. Now there are two sets of test scores, one ensuing from administration prior to hiring, and the other taken after 30 days of employment. The results from these two testings are compared in Table II.

A comparison of the two sets of scores reveals great differences in the test per-

TABLE II

Applicant-Employee	Pre-Employment Score	Ranking	30 Day Score	Ranking	Total Sales For First 30 Days
Jones	125	2	50	9	630
Smith	80	4	200	1	400
Brown	22	10	80	5	110
White	47	8	101	3	235
Peters	216	1	79	6	1080
Curry	99	3	69	7	480
French	50	7	91	4	250
Black	72	5	192	2	360
Lawrence	31	9	67	8	151
Tyler	65	6	32	10	325

First Testing Correlation
r = +1.00

Second Testing Correlation
r = +.03

formance of individuals from the first test to the second test. Furthermore, there seems to be no particular direction to the change. Employee Peters, for instance, achieved the highest test score on the first testing, and was therefore ranked number 1; on the re-testing he slipped to sixth position in the ranking. Similarly, all the employees changed ranks within the group from the first testing to the second testing.

A correlation computed between initial test scores and 30-day sales output you will recall was a sizzling +1.00. When a correlation is computed between the 30-day test scores and sales output, a somewhat dismal +.03 coefficient is obtained.

Now the poor sales director is thoroughly confused. If he had employed results from the first test administration to select salesmen, precisely the very best could have been chosen. Results from the second testing did not correlate to any appreciable extent with sales output. Use of these results to select salesmen would have been a disaster. Which result is the correct one? Is the test highly valid as the first correlation suggests; or is it a dud — a conclusion drawn from the second correlation?

The sales director decides more data are needed. He re-administers the test a third, fourth, and fifth time to the same 10 employees. Each time new rankings occur; a new highest and a new lowest scorer emerge. Correlations are computed as each array of scores is related to 30-day sales output. Their correlations are given below in Table III.

TABLE III

Correlations From Multiple Administrations

1	2	3	4	5
+1.00	+.03	-.72	+.01	-.15

Each of these correlations could be taken to define validity for the test. But which one is the correct one? What is the "true" validity? The sales director may repeatedly ask himself these questions as his dreams of the company vice-presidency fade, but he will find no answer.

RELIABILITY — CONSISTENCY OF RANKINGS

Reliability can be viewed in one sense as the tendency for scores associated with a particular individual to consistently retain the same rank or position relative to other individuals' scores in repeated administrations of the test. To the extent that there are changes in rankings, unreliability is in evidence.

Defined in this fashion, the importance of reliability to the realistic business of test construction should be evident from the example above. Simply speaking, *without adequate reliability, validity cannot exist.* Unreliability in a test, for all practical purposes, means invalidity — and therefore, a worthless test.

The strongest statement the sales director can make about his test is that it sometimes appears to be related to sales output, and sometimes it does not. When dollars and cents decisions are made in terms of who is to be selected and who is to be rejected from among those applying for employment, a definite statement of the test's validity must be available before it can reasonably be employed.

CONSISTENCY OF RANKS AND OTHER MEASUREMENT

It would be strange, as well as unfortunate, if the definition of reliability, as discrepancy in rankings in repeated measurements, applied only to psychological tests and measurements. Even the ruler must dance to the tune of reliability.

Suppose the social director for a fraternity is arranging an exchange dinner with a sorority. Since the plan is to pair-off each sorority member with a suitable dinner companion from among his brother stalwarts, the question of compatible heights is again a strong concern. His counterpart within the sorority has conveniently furnished him with a list of names along with the heights of her sorority sisters. A simple approach to the problem is to assign the tallest boy to the tallest girl, the second tallest boy to the second tallest girl, etc.

But first, the social director must have the heights of his fraternity brothers. A pledge is assigned the job of securing the heights of each fraternity member. Unfortunately, lowly pledges are not allowed to initiate conversations with active members. The pledge must therefore measure each member and record his height: he borrows a ruler from his roommate's desk for this purpose. His roommate, however, is an amateur magician and unbeknownst to the pledge, has been rehearsing with this particular ruler for the upcoming fraternity talent show. The ruler is made of a special substance greatly affected by temperature. Within a temperature drop of even a few degrees, the ruler may contract so strongly that a foot compares with 6 inches on other rulers. In slightly warmer areas, the ruler may stretch to the point that one foot is two or even three feet by normal standards.

Oblivious to the vicissitudes of the ruler, the pledge makes his measurements and returns his recorded heights to the social director, who in turn matches couples for dinner. But the dinner is a disaster since it finds such individuals as Stella String, Campus Armpit Champion, paired with young Billie Wee, heist man on the much touted pantie-raid team.

Measurements made by the pledge have not been valid. They have not related to measurements of the girls' heights made by the sorority social director. There are several possible causes for this lack of validity:

1. Measurements made with one ruler may, by nature, never relate to measurements made with other rulers. Naturally, we must reject this possibility. Historically speaking, the fact that measurements made with one ruler at one time and place will relate to measurements made with another ruler at another time and place, is the practical feature and convenience which both initiated and retained the use of the ruler as a measuring device.

2. The pledge's measurements are unreliable. This means that if the pledge were to measure all of his fraternity brothers again, he would find changes in their relative heights. These discrepancies in variations from one measurement to another, define unreliability as we have presented it. Strictly speaking, it is more accurate to say that the pledge's measurements were unreliable rather than the ruler he used. By dint of our privileged position, we know that the ruler was in fact a "trick" ruler, and that measurement made by persons other than the pledge would also be unreliable.

Some individuals, taking cognizance of the effect of temperature on the particular ruler could, by either controlling the temperature or adjusting the measurements for temperature effects, come up with measurements which, in fact, are reliable and valid. By the same token, even the most standard ruler can be employed in such a manner that it will yield unreliable measurements. When a ruler, psychological test, micrometer, or other measuring device is spoken of as being reliable or unreliable, there is an implicit reference to a standard method and procedure of applying the device or instrument.

3. It is possible that the measurements supplied by the sorority were unreliable or that both sets of measurements were unreliable. Unreliability of test measures, criterion measures, or both, will always destroy validity. In our example, unreliability was assumed to have resulted because of the "trick" ruler. Of course, there could have been other sources of unreliability.

CHANGES WHICH DO NOT AFFECT RANKS

In an earlier example in this chapter, a sales director discovered unreliability which was defined as a change in ranks from an initial testing to a subsequent re-administration

Once again let us presume that the sales director administers his tentative test to the same 10 applicants prior to hiring them, and then re-administers the test after 30 days of employment. The results of these two testings, along with the total sales made by each individual during the first 30 days, are presented in Table IV.

Inspection of pre-employment and 30-day scores in Table IV shows that all applicants scored 50 points higher after 30 days on the job than they had scored on the test before beginning work. Evidently, work experiences increased their ability to answer test questions. But what effect will this have on the relationship between test score and sales performance? Surely, this will affect at least the magnitude of the correlation coefficient.

Reference to Table IV indicates that there is no change in the relationship between test scores and sales performance. Correlations coupled from either pre-employment scores or test scores after 30 days on the job are identical. Both are a perfect +1.00.

It is a feature of correlation, and of prediction in general, that increasing or decreasing measures by a *constant* amount will do nothing to change the relationship between the two sets of observations. It will, therefore, not alter the precision with which prediction can be made from one to the other.

Suppose that instead of using the sales output of the first 30 days as a criterion, we would use the sales output from the second 30 days, or for that matter the third, fifth, or even twentieth 30-day-period. Provided that sales for all individuals increased by a constant amount, there would be no change in the observed correlation coefficient, and, therefore, no change in predictability from one set of observations to the other. This feature is summarized in Table V.

The relationship pointed up in this section may be summarized by saying that any changes in the magnitude of either test scores or criteria which do not affect the rankings on either set of observations will not change our ability to predict the outcome of one from knowledge of the other. (NOTE: Influences which do not affect all measures in either set of observations equally will change the magnitude of the coefficient of correlation when the Pearson Product Moment Coefficient of Correlation is employed to compute it. However, provided that ranks do not change, other statistical treatment will yield correlations equivalent to those seen in this instance.)

Regardless of its effect on prediction, measurement situation which yields a numerical statement of "six inches" on the first occasion a block of steel is measured, and on subsequent measurements yields lengths double or triple the original one might be questioned. If such an occasion did arise, we would tend to look for a "trick" ruler or convince ourselves that an illusion had distorted observation. Since we have never heard of steel blocks which suddenly grow or change shapes, we look to the measuring device, or to the procedures of the individual making the measurements for the source of error.

Where psychological measurement is concerned, changes in the objects to which measuring devices are applied are commonly experienced. It is the behavior of the individual in relation to the test instructions and devices which evoke the measurement statements. Human behavior is a dynamic entity. It changes constantly as new experiences leave their effect on the organism. The individual does not lie dormant in a cloistered pumpkin shell between his first exposure to a psychological test and a subsequent one. New learning, biological changes, or alterations in structure due to injury or maturation can drastically modify modes of responding, and thus results, from one testing to another.

Actually, changes in behavior due to new experience, and the stability of a block of steel during the same span of time, is more a relative difference in the nature of the two entities which are measured than it is an absolute one.

Some changes the steel might undergo such as temperatures, bombardment with high energy particles, (such as those emitted by a lazer), or long exposure to chemicals (even those in the air) can drastically modify the way the block of steel "behaves" from one exposure to the ruler, or micrometer, to a subsequent one. Obviously, the conditions which alter the form of human behavior are more numerous, and act faster, than those which alter the "behavior" of steel.

One feature of psychological measurement which is almost non-existent in physical measurement occurs when an initial measurement procedure changes the entity being measured. Outcomes of subsequent measurement differ from the initial outcomes because some aspect of the object is changed by the measurement procedure itself. A length of lumber may decay after years, or be consumed in a holocaust, but one hardly expects it to warp, crumble, or shrink because a ruler has been laid on it. Individuals, on the other hand, may remember test questions to which they later learn the answer. As a result, in a future testing they may be more proficient at problems or puzzles they encounter, due to practice during an earlier experience with the testing procedure; or they may simply move faster and more efficiently through the test on a repeated exposure because of a general familiarity gained from the first testing.

In these instances a previous experience with a specific test will change outcomes on the same test administered at a later time. This experience is peculiar to a particular test and would not be expected to change outcomes on other tests administered subsequently.

There are, however, data to support the belief that repeated exposure to a variety of tests will change performance on all subsequent tests, including those which have not been previously encountered. Through repeated testing experiences individuals learn to take tests. It is similar to what Harlow has demonstrated

TABLE IV

Applicant	Pre-Employment Score	30 Day Score	Total Sales For First 30 Days
Jones	125	175	630
Smith	80	130	400
Brown	22	72	110
White	47	97	235
Peters	216	266	1080
Curry	98	148	480
French	50	100	250
Black	72	122	360
Lawrence	31	81	151
Tyler	65	115	325

Correlation between pre-employment scores and sales r = +1.00
Correlation between 30 day scores and sales r = +1.00

TABLE V

Test Score	Sales Output (Validation Criterion)	Correlation
Increase or decrease all scores by a constant amount	Retain	+1.00
Retain	Increase or decrease sales by a constant amount	+1.00
Increase or decrease both by same amount		+1.00
Increase or decrease one by one amount Increase or decrease other by another amount		+1.00

with primates who become increasingly efficient at solving new problems of all sorts because of an ongoing intensive exposure to a variety of problems, and because of the similarities in the problem-solving behaviors required.

Both test-wiseness, and experience with a specific test, will produce unreliability, and nullify any possibility of relating test results to criterion measures. Consider the following situation.

CRISIS IN CARPATHIA

A government builds a new plant somewhere in Carpathian mountains near that geographical area known as Transylvania. The employees for the plant will be recruited from among the local inhabitants, a hardy and robust people, who are generally excellent workers. Problems soon arise, however, because werewolves and vampires (who are masters of disguise and can make themselves appear normal) are inadvertently hired. If vampires are hired for the day shift they do nothing but lie around all day. If hired on the night shift, they literally weaken the work force due to their peculiar interpersonal relationships with other employees. Werewolves are excellent workers sometimes, but are given to occasional periods of moodiness and depression during which they are highly irascible, uncooperative, and bad for plant morale.

A psychological test is soon constructed in an attempt to screen out the vampires and werewolves. The new test is administered to five applicants who are hired and put to work immediately. Job performance data are collected for 30 days. At the end of this period test results are correlated with performance data. This relationship is presented in Table VI.

It is obvious from the performance data that neither Wolfman nor Dracula have been crackerjack employees. They will have to be dismissed. Their low test scores are consistent with low performance ratings. If test scores had been used to screen new applicants, they would not have been hired in the first place. The fact that test scores are predictive of work performance is, of course, suggested by the extremely high correlation (r = +.964).

The test appears to be a valuable device in the selection of future employees. But it must be further validated with new applicants. As already indicated, vampires and werewolves are masters of disguise, so the next time that word goes out that the plant is accepting new applicants, they change their names, and appearance, and re-apply. They are subsequently given the same test they have taken before, and are hired along with five other applicants. Test scores for Wolfman and Dracula from the first testing and second testing are presented in Table VII, as well as performance data from the first 30 day period and the 30 days post-re-hirement. Test results and 30 day performance data from the five new applicants are offered for comparison.

Table VII shows that both Dracula and Wolfman scored 100 points higher on the second testing than they did on the first. Presumably, this increase was due either to experience manifesting itself between testings or to the fact that they had learned how to score higher on tests due to the repetition of test-taking-behavior. Since the change was the same for both men there is a tendency to accept the premise that the increase in the score was a direct result of practice, and learning, transpiring *during the first testing.*

A comparison between performance data for the first 30 days of work, and the second 30 days, does not reveal any change. The increase in the test score therefore is not associated with a comparable increase in work performance. Wolfman and Dracula are certainly no more acceptable as employees than they were previously. Again they must be dismissed.

Test scores are still correlated with work performance (r = +.83) but the magnitude of the correlation is reduced. This means that the test score is less predictive of success. Thus, the test is less valid.

Provided that each exposure to the test results in increased scores, the test constructor has real problems with his test. It is easy to see that if Wolfman and Dracula, along with their brothers, cousins, and uncles, continue to change their identity and re-take the selection test, test scores will correlate with work performance in an ever-lessening fashion. Test scores will be worthless as predictors of work performance and the test will be useless.

Previously it was pointed out that influences which changed all test scores (or criterion scores for that matter) by a constant factor would not change the value of the coefficient of correlation, or change the possibility for prediction.

In practice, however, it is often impossible to control or discover the incidence of previous experience each applicant or testing subject has had with a particular test. This means that differences in test scores among individuals may be more related to the number of previous exposures an individual has had to the test, than they are to the validating criterion. If administered to individuals who were new to the test, or who had had the same number of previous exposures to the test; and further, if each exposure had the same effect on each individual's test performance, correlations between test scores and validating criteria may be extremely high. Differential test experience among a group of subjects thus accounts for low validity.

When "test-wiseness" is considered, the practical impossibility of control, or assessment, of test experience is more apparent. In contemporary America, commercial, educational, and public interests bombard citizens, from the cradle on, with tests of all descriptions and nature. This results in a citizenry of vast individual differences in terms of test experience.

TABLE VI

Applicant	Rank	Test Score	Performance Data	Rank
Jones, Samuel Park	1	206	100	1
Smith, Henry Lee	2	190	98	2
Brown, Frank Wm.	3	160	90	3
Dracula, Noel Account	4	62	5	4
Wolfman, Harris Swarthy	5	-4	3	5

r = + .964

TABLE VII

Applicant	Rank	Test Score 1	Test Score 2	Perform. Data 1	Perform. Data 2	Rank
Lowery, James Dee	1	—	200	—	98	1
White, William Henry	2	—	192	—	91	2
Perkins, James Fred	3	--	181	—	87	3
Curtiss, Frances S.	4	—	170	—	81	4
Jones, Joseph Philip (alias Dracula, N. A.)	5	62	162	5	5	6
Putney, Brian Curry	6	—	160	—	78	5
Smith, Peter Samuel (alias Wolfman, H. W.)	7	-4	104	3	3	7

r = + .83

RELIABILITY — TOWARD A DEFINITION

Unreliability was described earlier as a change in ranks in a group of test scores associated with specific individuals, from one test administration to a subsequent one.

Later it was stressed that any influence which affected all test scores in a group, in an equal fashion, would not change the nature or magnitude of a correlation between test scores and a validating criterion.

Finally, it was pointed out that due to the practical impossibility of controlling either specific or general past test exposure, and the similar impossibility of controlling experiences transpiring between the first testing and a subsequent one, any influences which would result in a change in the test score, associated with an individual, from one testing to another will result in lowered validity. Measuring instruments and procedures generating results or scores which are affected by the conditions mentioned above are unreliable.

One way to define reliability is to say simply that it is the tendency for the same test results to consistently be in evidence for a particular individual on repeated administrations of a test.

This is, of course, what we mean when we refer to the reliability of a ruler, or other device, used in physical measurement. Repeated measurements of a block of steel should yield results which are quite similar. If this condition is satisfied, certainly a group of steel bars of differing lengths will retain the same variations on different measuring occasions. The ruler would be said to be reliable. Measurements made with it will be valid because we can predict that a board cut to a specification defined by this ruler will later fit into a space for which it was cut.

However, where psychological measurement is concerned, there is an exception. We cannot, with qualification, speak of reliability as correspondence in test scores on repeated administrations of a test. Consider the following situation. In Table VIII the same five individuals are exposed to two separate administrations of a vocabulary test. The score on the vocabulary test is correlated with intelligence quotients (I.Q.) as defined by a standard intelligence test.

TABLE VIII

Subject	1st Vocabulary Test	Rank	2nd Vocabulary Test	Rank	1st I.Q. Test	2nd I.Q. Test
Jones	100	1	100	1	135	135
Smith	91	2	91	5	128	128
White	86	3	98	3	120	130
Curry	70	4	99	2	114	128
Young	62	5	97	4	107	130

If attention is focused on the vocabulary test scores from the two administrations there is a definite change in the scores of White, Curry, and Young. These three subjects all score considerably higher on the second test than they did on the first. This lack of consistency in results of repeated testings defines unreliability as it has recently been presented.

We must not give up here, however. Inspection of the outcomes of the two inteligence tests administered reveals that as vocabulary scores increased for White, Curry, and Young from the first to second testing, intelligence test scores also increased. Thus the relationship between vocabulary scores and intelligence scores has not changed. Prediction of intelligence from vocabulary scores (or vice versa) remains constant. The apparent unreliability has not affected the validity of the vocabulary test as a prediction of intelligence.

The situation is peculiar to psychological tests and measurements. To illustrate, imagine the following.

You wish to place a shelf across one entire wall of your den. First you measure the wall and find it to be 10 feet long. Leaving the house you go to the garage where your lumber is stored and find a board approximately 10 feet in length. Measurement done on the spot reveals that the board is exactly 10 feet long. It should be just right.

Returning to your den you measure the board once again. Impossible! Now the board is 20 feet long — twice as long as measurement indicated in the garage. Just as you are about to cut the board in two, you decide to double check your measurement of the wall. This time the wall measures 20 feet. This measurement too has doubled. Now you take time to contemplate whether you should call the psychiatrist or the carpenter.

Hopefully, this situation is not one the weekend carpenter must suffer. Typically, boards and walls do not behave in this fashion. Psychological tests, sad to say, do contain this element of rapid change as well as other peculiarities we have mentioned. It is, therefore, important that we qualify the definition of reliability in a manner which will be somewhat different from the concept as it occurs in other types of measurement.

Let us advance the following:

Within a given psychological test, reliability is the tendency for comparable individuals, or the same individual on repeated testings, tested under standard conditions, to achieve identical test scores.

This rather loose definition seems to suit our purposes for the moment. It takes into account the various points we have discussed thus far.

MEASUREMENT'S STEP-CHILD

Throughout preceding pages the importance of reliability as a necessary condition for validity has been stressed. Perhaps it is time to confuse the reader by saying that reliability is, *on a theoretical level*, quite beside the point. Below are some salient conclusions which should be made before we discuss the practical procedures employed to assess reliability.

1. The ultimate worth of a psychological test is strictly dependent on its validity. If scores generated by a test can be shown to consistently relate to an appropriate validating criterion, reliability need never be mentioned.

2. A test which produces scores of highest consistency, and by all definitions is extremely reliable, may be entirely worthless for particular purposes. Reliability in no way insures validity.

3. Although one need not necessarily concern himself with reliability when validating a test (#1 above), in practice it is most wise to do so. Inasmuch as a test produces unreliable scores, it is neither possible to validate nor is it reasonable to attempt to do so.

METHODS FOR ASSESSING RELIABILITY

Because test construction is a realistic attempt to apply principles of behavioral science to problems of everyday importance; time, effort, and expense cannot be disregarded.

The most practical place to begin with the validation of a test is first to assess reliability. There are several methods which are typically employed to assess reliability. Most of them involve a statistical definition of consistency, usually employing the coefficient of correlation.

TEST, RE-TEST RELIABILITY

The most common and straight-forward technique employed in the assessment of reliability is the test, re-test technique. Once test items or problems are selected and solidified into an integrated package assuming the form of a psychological test, a representative group of subjects are chosen and the test is administered to them.

Sometime thereafter, within a week or two, the test is re-administered to the same group of subjects. Scores from the first testing are compared to scores from the second testing. If the test is reliable, we would expect each individual to score the same on both testings, thus retaining his relative position in the group. Inasmuch as scores are consistent the coefficient of correlation will approach +1.00. Differences between the two scores will tend to reduce the size of the coefficient.

Some factors inherent in the test, retest method act to reduce the size of the coefficient of correlation, and thus, define a lower level of reliability for the same test than other methods.

In some cases, particularly if the time between test and re-test is short, experience during the first testing may influence scores on the second testing. If the subject remembers answers or otherwise learns specific responses, scores from the second testing will be higher than initial test scores.

A long delay between testings offers the possibility for any number of influences to intervene, and creates a lack of stability in scores. In addition, the task of gathering a sometimes large group of subjects on two separate occasions is expensive and inconvenient.

The fact that individuals are subjected to the very same test on two occasions insures that test scores are in response to identical item populations. If the difficulties mentioned above can be eliminated, differences between scores are indicative of a poorly constructed test. Unreliability is most likely attributable to features such as unique test instructions, inconsistent scoring policy and procedures, or other aspects inherent in the test.

EQUIVALENT FORMS

Often when a test is constructed it is desirable to construct not one, but two tests which are designed to serve the same purpose. These counterparts are really separate tests which are designed to be used interchangeably — thus they are termed *equivalent forms.* Although covering the same content, questions and test items are different in two forms.

Equivalent forms can be employed in a re-test procedure in much the same manner in which a single form is used in the test, re-test procedure. Subjects are administered the first form, and shortly thereafter, the second form is given. Then a correlation is computed between scores on the two forms.

Since items are different on the two forms there is little likelihood that a "practice effect", or other specific learning from the first test, will influence outcomes on the second. This means that the second test can be given soon after the first, and the group does not have to be brought back a second time for testing.

The major drawback of reliability estimated in this manner, is the possibility that the two forms are not truly "equivalent". If one form is quite different in even minor respects from the other, certain individuals may attain divergent results on the two forms. In general, reliability defined by equivalent forms is lower than that defined by other methods.

Most certainly, an equivalent form would not be constructed for the sole purpose of assessing reliability. More often, equivalent forms come about because of other practical advantages to having two "separate but equal" tests. Where equivalent forms do exist, however, they allow for a convenient, and ready, method of reliability assessment.

SPLIT-HALF RELIABILITY

Tests having a substantial number of items can be split into two homogeneous halves, and scores for each individual from the first half correlated with scores from the second half. The logic of this procedure is not substantially different from that employed in the equivalent forms method. In a sense, one half of the test has one equivalent form in the other half inasmuch as it offers what is presumably the same content, although it doesn't have items in common.

Obviously, the split-half method requires that the test be divided in a fashion which yields two equivalent halves. Items are arranged in order of increasing difficulty and odd-numbered items assigned to one half while even-numbered items are assigned to the other.

The split-half mehtod is inexpensive and convenient because it requires the administration of only one test to a group of subjects. There is neither a problem of "practice effects" as with the test-re-test method, nor is there opportunity for intervening influences between testings since there is only one administration involved.

Due to the test itself, the split-half gives the purest picture of unreliability. Differences in mood, fatigue, health, or motivation which may occur from one testing to another are not in evidence. These influences are constant in split-half determinations because only one testing is involved. Differences in scores from one half to the other are more likely attributable to features of the test rather than to extraneous variables.

PRACTICAL DECISIONS BASED ON RELIABILITY

All methods of defining reliability have various advantages and disadvantages. In specific applications one method will yield a larger correlation than another. In evaluating these methods, three things must be remembered.

1. They do not guess, or estimate, "true" reliability; rather, reliability is defined by these procedures. While a broad definition of reliability such as presented earlier in this chapter is of some advantage in delineating general principles, reliability, as a workable tool in psychological tests, is defined by the specific procedures employed in a given instance. In most cases, reliability boils down to a coefficient of correlation. The relevance of this correlation depends on a variety of factors concerning the purpose of the test and its ultimate validation.

2. The method of assessing reliability which yields the highest correlation is not necessarily the best index of its possible validity. A consideration of reliability is only a means to an end. A coefficient of correlation of acceptable magnitude is merely a "green light" which tells the test constructor that there is at least hope of finding validity. Without reliability there is absolutely no hope of validity.

The choice of a method to assess reliability is a practical one. Such factors as the nature of the test problems or items, the length of the test, the likelihood of practice effects, and the population of of individuals for whom the test was designed must all be considered in the choice of techniques to define reliability.

3. The methods of assessing reliability which we have discussed so far are by no means the only ones. Methods employing analysis of variance, factor analysis, and other statistical techniques are beyond the scope of this book. These and other techniques are necessary tools in the armament of the professional concerned with the construction and application of psychological tests.

RELIABILITY — A FINAL WORD

In the concepts of reliability introduced in this chapter, as with the content of previous chapters, there has been an attempt to view the material in a theoretical and philosophical manner. It is important that this attempt be made, because we must seek to relate that which we teach to wider and more general features of human existence.

But do not be distracted by the philosophical filigree! Remember, a psychological test is a practical innovation. Ultimately, it must predict. Like all tools, it is used by people to do a job. If it does not fulfill the task, it is pointless.

In the following chapter we shall see how a lone Eskimo harnesses the power of behavioral psychology to tackle problems of romance in the frozen North. We shall eavesdrop as he goes through the various steps in test construction which terminate in a test instrument which "nose" true survival value.

Surely those of you who have traveled

this far cannot miss the thrilling adventures which await you in the spine-chilling climax of this book!

REFERENCES

Anastasi, A. *Psychological Tests.* New York: Macmillan, 1968. pp. 99, 127-131, 78-80, 71-72, 92-94, 72-76, 105-111, 80-84.

Freeman, F. S. *Theory and Practice of Psychological Testing.* New York: Rinehart & Winston, 1962. pp. 100-104, 69-72, 73-77, 81-86, 66-69, 77-79, 88-91, 97-98, 72-73.

McNemar, Q. *Psychological Statistics.* New York: Wiley, 1969. pp. 163-171, 173-176.

Young, R. K. & Veldman, D. J. *Introductory Statistics for the Behavioral Sciences.* New York: Holt, Rinehart, & Winston, 1965. pp. 353-354, 356-359.

CHAPTER 10

Constructing a Test

STEPS IN TEST CONSTRUCTION

Like mariners home from a perilous voyage, we must now attempt to make sense of, and evaluate, our meanderings through previous chapters. Having, presumably, dispatched basic principles, and unsnared the reader from the confusion and vagueness engendered by his exposure to other "experts" on the topic, we are now clear to go full speed ahead in demonstrating how these principles are applied in an exercise of test construction.

There are several basic steps in the construction of a psychological test. They include:

1. Item selection and content consideration.

2. Definition of the population and sample selection.

3. Standardization, including reliability and validity assessment.

4. Establishment of test norms.

5. Use of the test for the purpose it was constructed.

NONOOKI AT 80° BELOW

It has been said that "necessity is the mother of invention". In the following example of test construction it is dire emergency which impels the hero of our piece to resort to the construction of a psychological test.

Our hero's name is Nonooki. He is a young Eskimo lad who has grown disenchanted with the variability he encounters in nose-rubbing techniques among the young girls he dates. After all, in the Land of the Midnight Sun, an evening lasts for months. Spending his time with an inept companion can get a bit tiresome. It would be convenient if some data about the girls' expertise in nose-rubbing were available before he asked them for a date. Nonooki elects to construct a test for just such a purpose.

SELECTION OF TEST ITEMS

Nonooki knows he must first decide on a series of test items. These items are questions, puzzles, or other forms of stimuli which will be employed to bring forth behavior from the test subject which can be expressed in numerical scores. In selecting test items, potential materials may be tried out informally on friends or relatives. For Nonooki it may have to be an unsuspecting team of huskies.

First of all, a test is much more than just items, questions, and — wait! Don't we already know what a test is? You will recall the definition introduced some chapters back.

A SET OF STANDARD STIMULI AND PROCEDURES, INCLUDING ENVIRONMENT, INSTRUCTIONS, PRESENTATION OF TASKS, AND SCORING CRITERIA, WHICH, WHEN APPLIED TO AN INDIVIDUAL, YIELD STATEMENTS WHICH CAN BE USEFULLY RELATED TO HIS FUTURE BEHAVIOR.

This definition points out the fact that items cannot be separated from:

1. The location where the test is given.

2. The individual who administers the test, and his general, and specific, behavior toward the subject.

3. Instructions, which should affect all subjects in an equal and standard manner, as well as control some general aspects of test performance.

Thus when Nonooki selects items for his test he must consider such things as how, and by whom, the items will be introduced, the atmosphere, including such things as igloo temperature, noise level, lighting, and general incidence of distraction. He must generate a set of instructions which is specific enough to eliminate all confusion as to what is to be done, but must use language which is not so technical that the ordinary "man on the tundra" will not understand it.

As for the items themselves, a few guidelines can be helpful to Nonooki.

1. Items should be chosen which allow for objective scoring. Inasmuch as there is confusion over whether an answer is right or wrong, or deserves 5 points credit as opposed to 8 or 6, there is unreliability. Naturally, this detracts from the probability of a high validity.

2. Whenever possible items should be chosen so that initial exposure to them will not affect performance on a subsequent administration. Items should be refractory to "practice effects".

3. As a general rule, the simpler the item in terms of steps to solution, method of marking answer, and complexity of apparatus and paraphernalia involved, the greater the possibility for reliability.

Aside from the considerations above, a decision must be reached on the number of items to be included. In Chapter 5 a generous amount of time was spent on topics related to this decision. Two points were most specific.

1. All things being equal, a test whose outcome statements, or result possibilities, allow for the widest range of differences to occur among test subjects will have the greatest possibility of attaining validity.

2. The test which actually yields the widest range of differences when administered to a group of subjects will have the greatest possibility of attaining validity.

Basically, there are two ways to gain the possibility for a greater number of differences to occur.

1. More test items can be added.

2. More possibilities for the definition of differences within a single item can be created.

Suppose a professor gives a five-question exam to his class of 100 students. A student either gets the question wrong or right — thus, scores can run from a maximum score of 5 to a minimum of zero. The tests are graded and results are as follows.

Score	Number of Students
5	20
4	20
3	20
2	20
1	20
0	0

The professor had intended to use his test to select the top student who would later be offered the prestigeous job of erasing the blackboard. From the results above not one, but 20, students would be offered the job.

It is simple to see that a test, such as the one above which allows for only six differences to occur among individuals, will probably always be of limited value.

Increasing the number of possible differences in the test could improve its chances of being serviceable. As suggested earlier, this could be done in either of two ways.

First, more questions could be added. Increasing the number from five to 100, or even 200, would greatly add to the structure of the test.

Second, individual questions can be broken down so that an answer can be other than totally right or totally wrong. Half, one-fourth, or one-eigth credit will have the same effect of increasing difference possibilities. Using one-half credit would increase the difference possibilities from 6 to 12; one-fourth credit would increase differences to 24; and one-eigth, to 48. The same 100 students re-tested on a five-question test with possibilities of one-fourth credit on each item might assume the following distribution.

Score	Number of Students
5	1
4¾	5
4½	7
4¼	7
4	5
3¾	5
3½	7
3¼	3
..	.
..	.
..	.
0	0
	N = 100

Obviously, the number of questions or items needed will depend on the nature of the questions and the form the results will take. If, as in the last distribution, subjects can score with a possibly high range of outcomes on a single question, fewer questions will be needed.

Whether the result of an increased number of questions, or a possible variation of the score within a single question, fewer than a total of 100 possible score differences are probably not enough to discriminate adequately among those taking the test. Beyond this guideline, the number of questions which should be employed will depend on feedback from trial use of the test on subjects.

Using this information, Nonooki selects test items which terminate in test results or scores of wide possible differences. Independent of the structure of the test items and the form of results they generate, is the expectation that the test questions will be relevant to the event the test constructor wishes to predict. In Nonooki's case, the important activity is sensual nose-rubbing.

There are several strategies which may assist a test constructor in making the decision that one test item is more likely to be relevant than another. By far the most common is *face validity*. We have already discussed this approach of letting common sense dictate item selection.

In other instances, a *psychological theory* may determine item content. In these instances, the validation procedure is often one similar to that described in Chapter 8 as *construct validation*. Preexisting assumptions about early childhood experiences and inability to participate in nose-rubbing may cause Nonooki to ask questions about the early personal history of his candidates. Certain facts of history revealed in this fashion could allow him to predict "frigidity" among Eskimo maidens who appear otherwise eligible for his attentions. The theory would guide him in the selection of test items which allow for prediction prior to committing himself for a long winter's evening.

EMPIRICAL VALIDATION

One method of item selection, strangely enough, requires no selection at all. It is referred to as *empirical validation*, and generally thought of as a validation method rather than an item-selection technique. It results, however, in a decision procedure which selects relevant items automatically.

Procedurally, it begins with those individuals who already exhibit the skill, talent, or behavior one would wish to discover in others. Because Nonooki is interested in discovering those who will be acceptable for nose-rubbing, when using this procedure, he would begin with those whose reputations already hold them to be among the most proficient. This group of young women with known gifts and talents are then subjected to a vast array of questions and tests of diverse nature and description. Literally, hundreds of questions, puzzles, conundrums, riddles, and tasks are stolen, borrowed, or begged from old tests, books, almanacs, or voodoo manuals, and then administered to the group expert at nose-rubbing. The content nature of the questions is of no concern. There is no attempt to judge the relevancy of the initial items. Topics of items may range from mechanics to early Byzantine art.

In addition to administering this vast population of tentative test items to a group of nose-rubbers of outstanding excellence, the same array of items are also administered to a group of individuals similar to the former group but of average, or even below average, nose-rubbing acumen.

There are several basic assumptions which tacitly underlie the empirical validation approach.

1. Individuals who are alike in exhibiting laudable nose-rubbing behavior are probably alike in other behavior patterns as well. These other commonalities of behavior may not be overtly related to nose-rubbing, and could be skills, habits, quirks, or idiosyncrasies of any description.

2. Expert nose-rubbers are obviously different from a normal, or unselected, group of individuals in nose-rubbing ability. They are also probably different from the unselected group in other systematic ways.

3. By administering a vast array of tentative test items to both groups, a subpopulation of items to which all expert nose-rubbers respond in one fashion, and all non-experts respond to in another, can be found.

4. Once identified, this subpopulation of items can be administered to a new group of subjects from whom those who are potential expert nose-rubbers will be selected. Those answering the questions and otherwise performing on the test items in the same fashion as expert nose-rubbers, will also eventually be like them in nose-rubbing ability.

Now the "automatic" nature of item selection, when the empirical validation method is employed, becomes apparent. Nonooki could let his test subjects select the items for him by their response to the hundreds of questions he hurls at them. He relaxes and ultimately chooses those items that all expert nose-rubbers respond to in another.

Much can be said to recommend the empirical validation approach. For one thing, once items are selected, validity of items has already been established. As a matter of fact, they were chosen because empirically they did relate to the validating criterion. On the other hand the empirical method involves brutish effort on the part of both the test constructor and his validating subjects. It is often an expensive and painstaking technique. It may continue on devouring countless items before enough discriminating items are found to enable them to compile a test of suitable length.

ACHIEVEMENT AND APTITUDE TESTS

A distinction made by those involved in psychological testing deals with the general nature of the behavior, or skill, a test is constructed to predict. The choice of test items is directly related to the conceptualization, by the test author, of that which is to be predicted.

An *achievement test* is one whose results are related to behaviors, skills, or responses which are typically acquired through learning. Educational tests of all kinds, most performance tests, practically any test in which test scores are intended to be related to changes in the individual occurring as a somewhat systematic product of experience, are achievement tests. Items on achievement tests are quite specific to a particular unit of experience. An achievement test on this chapter would involve the questions on content included here; and scores should correlate with the extent of your observation as you flip through these pages.

Aptitude tests generate results which are not as clearly related to specific units of experience. They seem to relate to more general skills, and broad behavior patterns whose acquisition cannot be tied directly to a chapter of a book, course of instruction, or any standard formal, or informal, learning situation. The behaviors to which scores from aptitude tests relate are learned in a variety of non-specific instances, and are more often linked in some integral fashion to the physical capabilities, or structures, or the individual. This emphasis on structural and constitutional factors, in terms of type of behavior an aptitude test reflects, has led many writers in psychological testing to define an aptitude as an "inherited pre-disposition". Hearing and vision tests are aptitude tests, and to a lesser extent, intelligence tests also qualify as such. Most tests employed to select individuals who are more likely to benefit from one kind of training rather than another, are aptitude tests of diverse types and have been employed widely in the armed services, in public agencies, particularly those having to do with rehabilitation or aspects of employment, and to a lesser extent, in industry.

In the past, a great deal of discussion has resulted from an attempt to classify one test or another as an aptitude, as opposed to an achievement, test. The intelligence test has often been the object of such discussions. Such debates seem pointless. While the distinction may help clarify important principles of psychological tests, it is impossible to find any skill or behavior which does not depend on inherited physiology and structure, or is not influenced by learning and experience of all types and origins.

STANDARDIZATION POPULATION

We have seen our hero, Nonooki, through a conglomeration of considerations, all pertaining to the selection of items for the psychological tests he hopes will save him from Artic disaster. By now, Nonooki has begun to wonder if it is worth it.

Nonetheless, he has come up with what appears to be a reasonable test. It is, at this point, still tentative. He has selected a group of items and generated test instructions, scoring-method and forms, and has specified all aspects of test administration clearly establishing a standard procedure. He has tried the test out on various individuals and after some modification is convinced that the test is ready to fly.

For our purposes, Nonooki's test is completed. The test must now be given in a more formal way to a standard group of subjects. First, he must demonstrate the range of performances he can expect in the future on his test. In order to find out he must select a sample of individuals whose performance as a group will give him a yardstick by which later performances of single individuals can be compared. This is, of course, the *standardization process.*

Obviously, if later, Nonooki plans to compare the performance of single individuals to that of the standardization subjects, he should select subjects for the standardization sample which are as similar as possible to the individuals on whom he will want to use the test later. If he wants to use the test to select from among ladies of the frigid North, he should use similar ladies in his standardization sample — malemutes, seals, and polar bears may have helped in selection and modification of items, but one could hardly use results from this sampling for standardization.

The standardization sample has to be *representative* of the population with whom the test will later be employed. We have discussed how samples become representative. In general, if the sample is extremely large relative to the population, it will almost always be representative; if selected randomly, a somewhat smaller sample can be used with the exception of representativeness prevailing. A smart test constructor will use a *random stratified sample* to insure representativeness. Nonooki does just that. He is careful to select subjects who proportionately, and collectively, are like the general population in terms of length of nose, susceptibility to head colds, nose temperature, etc.

After selecting the standardization sample so that it will be representative, the test is administered to this sample, thus completing the standardization procedure. Scores from this testing can be arranged in a frequency distribution, and scores summarized and described statistically. Characteristics, descriptive of the performance of the standardization sample on the test, comprise the test norms. An individual being tested later with this will be compared to the standardization group in terms of his performance relative to the mean, standard deviation, and specific performances of individuals in the standardization group.

Nonooki can say a great deal about any young lady to whom he subsequently gives the test. He can compare her performance to that of the standardization group and determine whether she is average, below par, or several degrees superior. But such comparisons are, at this point, somewhat meaningless, because they would pertain only to performance on the test, since no relationship between performance on Nonooki's test and dating behavior has yet been established. Nonooki has only begun the somewhat tedious process of test construction!

Selecting test items which are workable, generating instructions, establishing both general and specific standard conditions, and finally giving the test to a representative sample, is, in reality, only the beginning of test construction. From this point on, Nonooki's test can be as glowing as the aurora borealis, or as dismally bleak as the bleakest blizzard. Whether or not the test will help Nonooki choose an adequate companion is not yet an answerable question.

RELIABILITY

In content selection it is inevitable that the test constructor should get some informal estimate of the consistency of his test. Since friends, or other accidental populations of limited size, are often the hapless victims of not one, but numerous, exposures to the test in one stage of completion or another, the test constructor has an opportunity to see test results from repeated testings on the same individuals.

Indeed, without some notion of the reliability of a tentative instrument it would be foolish to proceed with selection and administration of the test to the standardization sample, or to begin serious thought of validation.

Ultimately, some formal reliability assessment procedure must be undertaken. This is particularly true if the test is to be published or otherwise widely circulated. One of the first questions a potential user will ask will be in regard to reliability of the instrument. Practically speaking, unless results ensuing from a test are reliable, the test is worthless.

So it is that Nonooki, having constructed a tentative test, and having selected a standardization sample and administered the test to them, comes face to face with the assessment of reliability.

Perhaps in other geographical locations Nonooki might use the test, re-test method. But dog sleds are slow and do not always run on schedule. Having assembled the standardization population once for an initial testing, Nonooki decides a subsequent testing in two or three weeks will be impractical. He prudently elects to employ the split-half method.

You will remember from Chapter 9 that the split-half method involves dividing the test items so that two equivalent halves result. This splitting of the test can occur after the test has been given as a single unit. The halves need not be given as separate and distinct test-halves.

Accordingly, Nonooki assigns items in a manner which insures equivalent halves. In doing so, he must be aware of *item difficulty*. Obviously, there would be little correspondence in scores from one half of the test to the other half if all the difficult items are assigned to the one, making the other a "snap".

Fortunately, having access to the data from the administration of the test to the standardization sample, Nonooki does not have to guess, or otherwise intuitively assess, the difficulty level of each item. He merely compiles a list of items and records the number of individuals who passed and failed each.

Those items failed by the most standardization subjects are most difficult, while those passed by the most subjects are the least difficult. By using this compilation of item difficulties, Nonooki easily assigns items so that each half of the split-half item populations contain an equal array of simple, moderate, and difficult items.

Once the equivalent split-halves are compiled, results for individuals on one half are correlated with results of the same individuals on the remaining half. When Nonooki completes his computations, a correlation of +.97 is achieved. This makes him jump into the air and click the heels of his muckalucks together. Such a correlation is indeed quite respectable and encouraging.

Actually, had Nonooki known of a practice applied almost standardly when split-half reliability is accessed, the correlation would have been still larger. A special correlation formula for the computation of the correlation from split-halves, called the *Spearman-Brown* formula, takes into consideration the fact that reducing the number of items in a test reduces the size of the correlation obtained. In the split-half technique each half is equivalent to the other, and to the test as a whole. It is the whole test for which reliability is being accessed. However, this estimate is made from a test which is essentially only half the size having only half as many items. This means that the correlation is smaller than it would be if two full length tests were employed. This discrepancy is corrected by the *Spearman-Brown* formula.

SOURCES OF UNRELIABILITY

When reliability is not in evidence there are a multitude o possible reasons. Some sources of unreliability are a function of the test material itself — others are independent or only tangentially related.

In Chapter 9 we spoke of some factors which could result in a low correlation when reliability was accessed. Practice effects and test-wiseness were two instances where unreliability resulted from previous test experience on the part of test subjects.

More often, unreliability results from poor test construction. Let us list a few features of a "sloppy" test which cause it to be unreliable.

1. Poor instructions to subjects. Subjects are confused or misled by ambiguous oral instructions given before the test by the administrator, or by instructions found in the written text of the test.

2. Ambiguous items. Unless the test constructor is interested in the subjects' ability to deal with ambiguity, there should not be ambiguity in the specification of the mode, or type, of response required of the subject.

3. Complicated or confusing test forms or response sheets. Unless the test constructor is concerned with the subjects' ability to discover the correct manner and location in which to record his responses, effort should be exerted to design the simplest and most foolproof method possible.

4. Variability in approaches to test paraphernalia. Tests requiring the use of mechanical devices or other paraphernalia often yield unreliable results because these devices are broken, or otherwise fail to operate in a standard fashion. Experience has shown that mechanical contrivances chosen must be practically indestructible before they can be practically employed in a test.

5. Uneven difficulty. In most cases it has been found a test should begin with simple items and become increasingly difficult as it progresses. This serves as a warm-up period for the subject, but, more importantly, reduces the probability that the subject will become stalled on difficult items. If this stalling occurs early, further items cannot be dealt with because of lack of time. Unless it's a "speed test", reliability will be highest when all subjects complete all items.

There are numerous other possible sources for unreliability, such as standardizing the test environment, test administration, and scoring. Variability occurring in the performance of the test administrators or their administration, will always contribute to unreliability.

Even after all the sources of unreliability that we have mentioned are eliminated, it may still be the case that individuals score differently on subsequent exposures. Such unreliability is unfortunate for it may mean hours of painstaking work and investment has gone for naught. Why such ultimate variability prevails is not a question we can answer here. Its origin becomes a problem to be dealt with through further research. We can only be sure it is not the work of an evil genie, mischievous leprechaun, or angry god. It is not a whim of nature, but rather a manifestation of relationships that our ignorance must transcend.

VALIDITY — HERE'S THE RUB

Hopefully, Nonooki's euphoria in finding reliability for his test will not leave him ignorant of the reality that the true test of his test lies ahead. As we have said before, reliability in no way insures that validity is what a psychological test is all about. Without validity a test is a pointless array of nosey questions.

There are many approaches to the assessment of validity. For the most part they have been discussed in Chapter 8. In this chapter we have introduced a new validation strategy called *empirical validation*. You will recall that it is a validation, and item selection, technique all rolled into one. Nonooki might well employ it.

Regardless of what validation procedure Nonooki will employ, any consideration of validation must start at the same place — the validating criterion. In Nonooki's case a validating criterion must be chosen which specifies nose-rubbing ability. If there were already an existing test for this ability, Nonooki could validate his test against this criterion. This is, of course, a form of *concurrent validation*. Alas, there is no such test. If there was such an animal, he could use it and save himself time and trouble. No, Nonooki is on his own.

In some cases validating criteria can be found in statistics or data compiled by other interests, or agencies, for other purposes. Nonooki is sadly out of luck here as well. No such data are available for nose-rubbing.

The very fact that he wishes to construct a test of this ability implies that judgments are already made in terms of discriminating good nose-rubbers from bad ones. Certainly Nonooki can tell the difference. Perhaps a group of his friends can do as well. Nonooki rushes to the Blubber Bar and Tap Room where he apprises his chronies of the important part they will play in the validating procedure. The plan he reveals to them involves the following steps.

1. A half dozen or so of Nonooki's companions (or other qualified connoisseurs of that sensuous past time) are enlisted to serve as judges or raters.

2. A sample of young ladies will be selected for validation purposes. It is important that the sample be representative of the population, and also be large enough so that conclusions drawn from it can be generalized to the population.

3. The test is administered to the vali-

dation sample chosen above who are later allowed to exhibit their nose-rubbing techniques to each of the judges who then independently assigns a score from 0 to 200 to each sample subject.

4. An estimate of inter-judge reliability is made. It must be ascertained that the judges are individually consistent, and comparable, in terms of tending to assign the same score to the same girl.

5. A correlation is computed between test scores and judges' ratings. This coefeficient of correlation will define validity. A substantial correlation means the test is highly valid and of obvious worth. Lower coefficients generally indicate that the test will only be of worth in certain specified situations.

SAMPLE SIZE AND INFERENCE FROM SAMPLE CORRELATIONS

Some of the points above deserve further comment. In number 2, for instance, mention is made of the size of the sample as a consideration. This is one feature of the coefficient of correlation we have not yet directly discussed, although we have dealt with the principles in the chapter on Statistical Inference (Chapter 7).

Correlations are computed on samples of subjects selected in such a fashion that they will be representative of the population of which they are a part. A correlation computed on data collected from sample subjects is descriptive of the degree of association found between two sets of scores or observations found in the sample. Naturally, all of this is done with the expectation that the same relationship, and therefore coefficient of correlation, would be obtained if computed for the entire population.

Thus, a feature of the sample correlation, other than its magnitude and direction, must be ascertained. It must also be *statistically significant.* If statistically significant, a coefficient of correlation found when sample data are statistically treated will also be descriptive of the population. The likelihood that such a correlation will be able to be generalized depends on the size of the sample, and is independent of attempts to stratify or otherwise make the sample representative through selection methods.

Specifically, our confidence that a coefficient achieved with a sample will also be true of the population increases proportionately as the square root of the sample size increases ($\sqrt{n}$). This feature of statistical inference was also discussed in Chapter 7.

Conventionally, both correlation and statistical significance of the correlation are computed. A high correlation which is not statistically significant is of questionable value since one cannot be sure such a correlation would also hold with the population. Naturally, it is the population which is of concern.

Statistical significance is not an all or nothing feature of a correlation. In particular instances statistical procedures which assess the statistical significance of a given coefficient of correlation allow probability statements to be made regarding the likelihood that the same magnitude of correlation would be found with the population as was found in the sample. The test constructor or researcher then decides to "take the chance" or to conclude it would be unwise to assume that the same correlation will hold for the population. Traditionally, most researchers will take a chance if the computed statistical significance is at the .05 level. This means that the test constructor is taking a 5 out of 100 chances that the correlation found for the sample is not also true of the population. More conservative test constructors may insist on a .01 (1 out of 100 chances) or even .001 (1 out of 1,000 chances) level of statistical significance. The decision as to what level will be accepted depends on a variety of practical factors concerned with the specific purpose of the test constructor or investigator.

VALIDATING CRITERION RELIABILITY

It was noted that inter-judge reliability must be ascertained before employing judges' ratings as a validating criterion. This is merely a reflection of a general concern which must always command attention. From the standpoint of the correlation computed between a set of test scores and validating criterion scores, unreliability in the validating criterion is no less damaging than unreliability in test scores. Both must be reliable before validity can obtain.

Nonooki's concern with inter-judge reliability is a manifestation of this general and necessary preoccupation. In ascertaining that judges are responding in a similar fashion toward the young ladies whose talents they must rate, Nonooki is keeping his chance for validity and a useful test alive.

Many test constructors are not so prudent. After spending great time and effort in constructing a test instrument, they attempt to validate the test with criterion measures which are unreliable. In some cases they may use validating criteria because they are convenient, and in so doing disregard the fact that such criteria are poorly related to the purpose for which the test was constructed.

Validity is for all practical purposes defined by a correlation computed between two sets of observations. It describes a relationship which is reciprocal. All of the considerations we have delineated regarding the language in which measurement outcomes are expressed, the importance of a wide range of outcomes, and of course, reliability, apply to scores or measurements of the validating criterion as much as they do to test results. Even the best constructed test cannot transcend poor validating measures.

PARADISE FOUND

The pay off for all of Nonooki's hard work rests precariously on the outcome of a single correlation. Judges' ratings for each individual will be correlated with test scores. A high correlation will mean the test will be effective in saving him from an endless and dark winter with a boring and uninspiring companion, and insure, instead, a blissful and joyous respite before the activity attendant on the spring thaw.

Not without trepidation, Nonooki completes the last of his computations. "Eureka!" he cries. The coefficient of correlation he finds is a smashing +.92. This means that by employing the test as a selection device he can choose thrilling companion after thrilling companion, far exceeding success gained by guessing, or using other arbitrary procedures employed previously. This is truly another in a series of man's conquest of the frozen Arctic!

But suppose Nonooki's test had not been valid? Suppose test scores were entirely worthless as predictors or indicators of nose-rubbing. Whatever will Nonooki do during the long frozen winter? Is there no way an invalid test can be of worth?

In such a case Nonooki had best learn to sleep because there is no therapy for an invalid test. It simply does not perform the job for which it was constructed. If reliable, scores created by the test may prove to relate to other events, outcomes, or criteria which were not considered when it was constructed. However, only further validation with these purposes in mind will begin to demonstrate this.

THE ESTABLISHMENT OF NORMS

But Nonooki's test is not a failure. It is highly valid and will, undoubtedly, serve its master with as much fidelity as any dog-team. Basically a humanitarian, Nonooki would not for an instant entertain the faintest notion of employing the test for his sole benefit; therefore, he will share the test and its magic with his friends, companions, and neighbors. Someday his test may become so much in demand he will publish it and circulate it widely.

In order to prepare his test for wider use, Nonooki must furnish complete information and materials to potential users. Naturally all test materials, including foolproof instructions for test administrations, will be included. In addition, thorough and specific details regarding validity and reliability assessment procedures and results will be demanded by those who intend to apply the test in other settings or locations.

At some time or another a test constructor may find a test becomes more valuable both to him and others if a set of *norms* are established. We have already alluded to norms in earlier pages, perhaps it is now time to specify this concept. A norm is a typical score (generally the mean or median) achieved by a particular sub-sample taken from the larger representative sample on which a test is standardized.

You will remember that Nonooki standardized his test on a large stratified sample. Individuals could be categorized in terms of age, geographical location, or scores on other criteria. Means or medians can be computed for each of these categories or sub-samples. Finally, an amalgamation of these categories and respective means or medians can be prepared in tabular form.

In looking at his standardization data, Nonooki finds no particular relationship between age and nose-rubbing ability. What he does find, however, is that by and large the farther from the North Pole his standardization subjects live, the higher the test score evidenced. This relationship encourages Nonooki to compile the following table.

TABLE 1

Norms For North Pole Proximity

Distance From North Pole (miles)	Mean Test Score
0 — 100	70
100 — 200	95
200 — 300	80
300 — 400	200
500 —1000	225

Mean For Total Standardization Sample M = 170

Table 1 furnishes valuable information to the test user. Without these norms one would be forced to use the mean from the total standardization group (M = 170) to estimate his score expectations from individuals tested. A seal hunter living within a 50-mile radius of the North Pole would be hard pressed to find a young lady who tested out with even an average score of 170. Reference to the table of norms demonstrates the origin of this difficulty. At this location the average is much lower (M = 70) than the overall mean. He would do well to lower his expectations and be less stringent in his selection requirements. Similarly, an acceptance of a score based on the overall average (M = 170) at a location 500 — 1000 miles from the North Pole would not be demanding enough. Certainly, the norms above suggest better talent is easily available.

FUTURE USE OF THE TEST

In some sense, a test is similar to wine — as it ages, it becomes more valuable and more useful. As its results are found to relate to other phenomena, scores, and systematic observations, worth of the test, both practically and theoretically, becomes more widely established. Generally, this occurs when the test is employed by individuals, other than the test constructor, for research purposes.

Frequently, the test constructor or publisher will extend the applicability and validity of the test by standardizing it on new populations and publishing new norms relevant to these populations.

However, not all changes tend to make the test more valuable. As new technological or environmental changes come about, new experiences change the behavior of individuals in the culture. These behavioral changes are reflected in modified test scores.

Previously established norms may become meaningless or misleading. In extreme instances the test may cease to be valid. It must then be discarded and a new instrument constructed.

Let us, like our hero Nonooki, enjoy the glory of his success. Knowing the nature of measurement, its relationship to everyday experience, and the specific proceudres for constructing a psychological test, Nonooki could do it all again if necessary.

Let us hope the reader can do as much.

REFERENCES

Anastasi, A. *Psychological Testing.* New York: Macmillan, 1968. pp. 12-16, 21-28, 71-79, 82-83, 86-89, 36-40, 104, 105-116, 127-131, 158-160, 167-172, 440-441.

Freeman, F. S. *Theory and Practice of Psychological Testing.* New York: Holt, Rinehart & Winston, 1962. pp. 13-15, 40, 67-77, 90-92, 112-114, 431-444.

Gronlund, N. E. *Constructing Achievement Tests.* Englewood Cliffs: Prentice-Hall, 1968. pp. 1-4.

McNemar, Q. *Psychological Statistics.* New York: Wiley, 1969. pp. 122-123.

Young, R. K. and Weldman, D. J. *Introductory Statistics for the Behavioral Sciences.* New York: Holt, Rinehart & Winston, 1965. pp. 167-168, 362-363.