
Tests

Small Samples, Large Consequences

W. Grant Dahlstrom

People make innumerable decisions daily about other people that affect their lives and careers. These decisions are inevitably fraught with errors of judgment that reflect ignorance, personal biases, or stereotypes on the part of the judges. Psychological tests can serve as a means to offset the impact of these distortions only if they are soundly constructed and responsibly applied. The requirements of dependable instruments are examined and the benefits from accurate test usage are reviewed.

For 'tis your thoughts that now must deck our kings, Carry them here and there, jumping o'er times, *Turning the accomplishments of many years into an hour-glass* [italics added].

—Shakespeare, Prologue, *Henry V*

Shakespeare bemoans the severe limitations imposed on him by a theatrical stage and the brevity of a play in his efforts to capture for his audience the vast sweep of English history and the tides of battle at Agincourt. We psychologists, too, are perilously constrained in our efforts to capture the range and depth of a person's lifetime of experiences and activities, skills and talents, successes and failures, all within the brief samples of behavior that we obtain in an hour or two by means of our psychological tests. Nevertheless, I have been repeatedly impressed by the level of success that we have been able to gain with our current instruments. It is quite impossible to document precisely just how well our psychological tests do work or the actual extent of their impact on the lives of people here and around the globe. In recent years we no longer examine these consequences quite so closely as some of the early psychologists did when psychological tests were first introduced. Nevertheless, in countries throughout the world, diagnoses and referrals, advice and guidance, and decisions on hiring and promotion, all involving a wide range of predictions, are being bolstered by the scores and patterns from psychological devices.

Fallibility of Human Judgments

People are poor judges of other people. In our dealings with one another we are subject to a variety of systematic and random errors. Prejudices and biases, based on race, gender, age, and even language, all too frequently distort our views of particular individuals. As Tallent (1992) pointed out, Aunt Fanny and Prosecuting Attorney ste-

reotypes may operate in our professional assessments—in the first instance, offering trivial commonalities that minimize important differences and, in the second, emphasizing a person's deficiencies while overlooking personal strengths and assets. Positive and negative halos, as well as the pervasive errors of central tendency, serve to cloud our judgments and evaluations.

In justification of their efforts to develop an objective and reliable means of assessing scholastic potential, Binet and Simon (1907) carried out a study of the dependability of the diagnoses of degree of mental retardation in children (i.e., l'idiotie [idiot], l'imbécilitéé [imbecile], et la débilité mentale [moron]) made by physicians examining young patients in several Paris hospitals:

We have made a methodical comparison between the admission certificates filled out for the same children within only a few days' interval by the doctors of Sainte-Anne, Bicêtre, the Salpêtrière, and Vaucluse. We have compared several hundreds of these certificates, and we think we may say without exaggeration that they looked as if they had been drawn by chance out of a sack. (p. 76)

This fallibility in the judgments made by humans about fellow humans is one of the primary reasons that psychological tests have been developed and applied in ever-increasing numbers over the past century. Certainly, Binet must have felt some deep pleasure in being able to demonstrate the replicability of the findings from even his brief intelligence scale on some child being admitted to one of the Paris hospitals, in sharp contrast to that devastating verdict he had published concerning the undependability of clinical judgments by staff psychiatrists. However, it is quite puzzling that we do not find anywhere in Binet's writings a summary report on a particular child whom he had tested and set on an appropriate course by using the results of his test.

Terman, on the other hand, in his extensive efforts to improve the brief instruments that Binet and Simon

Editor's note. Articles based on APA award addresses that appear in the *American Psychologist* are scholarly articles by distinguished contributors to the field. As such, they are given special consideration in the *American Psychologist's* editorial selection process.

This article was originally presented as part of a Distinguished Professional Contributions award address at the 100th Annual Convention of the American Psychological Association in Washington DC, in August 1992.

Author's note. Correspondence concerning this article should be addressed to W. Grant Dahlstrom, Department of Psychology, CB#3270 Davie Hall, University of North Carolina, Chapel Hill, NC 27599-3270.

had published, took repeated pride in his "real-world" successes. For example, he reported in great detail the facts in the trial of Alberto Flores (Terman, 1918), a 17-year-old Hispanic American facing the death penalty on a charge of sexual assault and murder in the death of a 6-year-old girl in Santa Barbara, California. Terman had intervened to help clarify questions concerning the mental competence of this defendant. First, his assistant, Miss Lamb, and then Terman himself tested Flores with the 1916 version of the Stanford-Binet (S-B; Terman, 1916). Both examiners obtained the same mental age (7 years, 6 months), although with slightly different patterns of successes and failures on the S-B subtests on two occasions, four days apart. Then Terman administered the Yerkes-Bridges Point Scale (Yerkes, Bridges, & Hardwick, 1915), a competitor of the Stanford-Binet, and obtained a mental age of 8 years. Both test results were consistent with the fact that Flores had dropped out of school in the third grade at the age of 15.

When the case came to trial, a forensic psychiatrist testified on behalf of the prosecution and stated his belief that Flores was a competent adult and should be held fully responsible for his actions. In support of this judgment, he offered evidence from a variety of informal and ad hoc tasks of his own devising that he had asked Flores to perform (e.g., recalling what he had for breakfast, counting tins of food in his cell, identifying objects in pictures from the morning's newspaper, and naming fellow prisoners). Terman, using the test results (plus detailed data on children, derived from the standardization of the Stanford-Binet), refuted each of the procedures used by the psychiatrist, clearly demonstrating that all of them could be done easily by children no older than seven or eight years of age.

In summary, Terman testified that Flores had the mind of a 7½-year-old and could not be held any more responsible for his actions than a child of that age. The jury believed Terman and rejected the death penalty for Alberto Flores. Terman could document his assessment based as it was on a standardized psychometric instrument; the psychiatrist had only his fallible judgment based on unreliable and invalid observations. The test data enabled Terman to break free of the error of central tendency that serves to render opinions about very bright individuals—that they are less intelligent than they really are—and about very dull individuals—that they are more capable than they actually are.

This salutary role of psychological tests in clarifying and documenting the bases for human judgment about other humans was also dramatically shown in the early 1930s in a report by Whitty and Jenkins (1935). Working in the public schools of south Chicago, they searched systematically for gifted Black children. Their method was similar to the one that Terman had used earlier in selecting subjects for his longitudinal study of very bright children in the California schools. That is, Whitty and Jenkins set out to examine with the Stanford-Binet successive pairs of children nominated by the teacher of each class: namely, the best (or model) student and the most intelligent pupil. In one

fifth-grade classroom, Barbara, the girl nominated as best student, 9 years and 4 months of age, earned a mental age of 17 years 5 months on the S-B. This was equivalent to an IQ of 187. However, because she was still passing subtests on the S-B at the highest level, no ceiling could be established. Obviously, her actual intelligence quotient would have been even higher. The girl nominated as the most intelligent in that classroom earned an IQ of 100 on the S-B, a full standard deviation above the class average but six standard deviations lower than Barbara. The error of central tendency was operating with a vengeance! Barbara was studied further by Whitty and Jenkins; she professed a strong interest in going to college to study chemistry. I wonder what kind of chemist she turned out to be?

My own experience can also offer evidence about the way that psychological test results can serve in a crucial way to break up a negative halo and improve one's clinical judgments. While I was still a trainee at the University of Minnesota Hospitals, a 15-year-old girl from northern Minnesota, whom I shall call Mary, was admitted to the inpatient psychiatric service on referral from her family doctor. Not particularly attractive, poorly dressed and disheveled, she was silent, withdrawn, and unresponsive to questioning by the psychiatric resident who admitted her. As his preliminary diagnosis he entered into her chart: mental retardation and possible schizophrenia.

At that time I had been reading about propf-schizophrenia (the condition of a schizophrenic disorder superimposed on intellectual retardation) and was quite surprised that we might have an actual instance of this combination of the two conditions on the psychiatric service. I made a note to test her soon and went to see the resident. He urged me to try to examine her as a dramatic example of poor protoplasm. However, before I could schedule her, the art therapist came and asked me if I had tested Mary. She had, with some difficulty, persuaded Mary to accompany the other patients to her service, but she had yet to hear her say a word. However, she had encouraged her to work on some drawings and was struck by her natural artistic talents. It was also her impression that Mary was not likely to be mentally retarded.

When I went to Mary's room, it was with some difficulty that I induced her to leave the chair in which she was huddled and walk with me down the corridor to the psychology office. Because she had yet to talk to anyone, I decided to start with the Object Assembly tasks of the Wechsler-Bellevue I (Wechsler, 1939). It worked! She assembled the pieces of the Mannikin quickly and accurately. I was even more heartened on the Profile task when she put the parts of the ear together outside the profile form and dropped them accurately into place. On the Hand task, holding up her own hand with a slight smile, she quickly saw the way that the elements of the Hand could be combined and put them together in rapid order.

Not wanting to press her into speaking, I skipped over the Picture Completion series. By the time that she started the Coding task, I was convinced that at least half of the resident's diagnostic impression was incorrect and was starting to doubt the other half as well. She listened

to the coding instructions and started the usual item-by-item substitutions. Then, with a sly grin, she shifted to entering all of the backward Ns for the 2s straight through the series, a way of beating the clock that some test subjects try. I was able to get her back on the appropriate coding sequence without any rupture in our growing rapport.

Although by now we had exchanged a few words, on this first occasion I decided to settle for the Performance IQ (about 115). But I knew that I now had enough information about her mental competence to overrule the admission decision not to have her take the Minnesota Multiphasic Personality Inventory (MMPI). I picked up a box of cards on the way back to her room, and before I left her, she was back in her chair busily sorting the items into true and false. When the MMPI results came back, there were only two scores appreciably elevated: Depression and Social Introversion. There was no evidence of any difficulty in completing the inventory, nor was there any suggestion of a schizophrenic process—no poor protoplasm, either.

I now believed that I had strong evidence to refute the other half of the admission diagnosis as well. I shared this opinion with the art therapist, and she continued her efforts to bring Mary out of her shell. Gradually the story unfolded: Her family was ruled by a father who was a religious tyrant. His wife was virtually helpless in defending the children, but one by one, the older girls in the family had managed to escape the home and come to the Twin Cities to establish new lives for themselves. Mary was the youngest, still living with her parents in an isolated community in northern Minnesota. However, just recently her mother had died, and Mary was then left alone to face the psychological abuse of her fanatical father, who irrationally blamed his children for his wife's death.

Mary herself was shy, introverted, and unassertive; her only escape was into a depressive withdrawal. Her sisters had learned of the changes in her behavior and had managed to get the family doctor to intervene on her behalf. The trip to the University of Minnesota Hospitals and the admission procedures had been extremely frightening to her, and for the first few days, she had been terrified of everyone. Only gradually was she reassured and secure enough to talk to us. Initially, her disheveled appearance, her silence, and her great fearfulness obviously generated an impression of serious psychopathology and, through the operation of a negative halo, led the resident to infer intellectual incompetence. His stereotype of poor protoplasm clouded his clinical judgment.

When the details of her circumstances were reviewed at the staff conference, the quality of the drawings that the art therapist displayed and the consistency of the test results that I had obtained turned the diagnostic formulation around. The evidence from the psychological tests bolstered my judgment, even as a trainee, in countering the diagnostic opinion of the psychiatric resident. The prognosis for Mary was now viewed more optimistically. With the help of a psychiatric social worker, her sisters were persuaded to keep her in Minneapolis, to help her get more therapy, and to support her in getting proper

training in art. I kept in touch with developments in Mary's career through the social worker and was delighted to learn later that Mary had obtained a position in one of the local art institutes. The last I heard of her was that she had been named the associate director of the institute and was gaining considerable local fame as an artist.

It is my belief that hundreds of thousands of scenarios of this sort could be documented on the constructive role that psychological tests now play in supporting and clarifying professional decisions about patients, clients, applicants, defendants, students, or employees in the United States and abroad. I have often pondered the bases for these achievements. What are the attributes of these procedures that engender them with the capacity to serve this way in offsetting the inherent errors and biases of human judgments about other humans? Halo effects, errors of central tendency, Aunt Fanny and Prosecuting Attorney stereotypes are ever present in situations in which one person makes crucial, career-determining decisions concerning some other person. Reliance on test findings serves to reduce the likelihood of the adverse effect of such misjudgment. Dawes, Faust, and Meehl (1989) have summarized the research bearing on the various elements that enter into the complex set of processes that comprise human judgment and have pointed out the central role that objective data such as test findings can play in the substitution of predictive formulae for much more fallible human integration of information about the individuals under study. Not only are specific formulae more likely to generate accurate decisions than are human judges using the same information sources, but computer-based interpretations of test data are also competitive in their accuracy in comparison with human interpretations of test findings (Eyde, Kowal, & Fishburne, 1991).

The Nature of Psychological Tests

As students at the University of Minnesota we learned to define a psychological test in the following terms: (a) standardized materials and procedures, (b) optimal motivation, (c) immediate recording, (d) objective scoring, (e) appropriate norms, and (f) established validity. (Any device failing to meet all of the criteria was not considered a test.)

Clearly, the procedures used by the forensic psychiatrist who examined Alberto Flores or those used by the psychiatric resident trying to evaluate Mary on her admission to the University of Minnesota Hospitals departed from these requirements in almost all respects. However, many other assessment devices also fail to meet these criteria. For example, were it not for the efforts by John Exner (1974, 1978) to improve the reliability of scoring, to establish some basic norms, and to document the interpretive accuracy of various scores and content material obtained by means of Hermann Rorschach's (1921) inkblots, the Rorschach method would fail to qualify as a psychological test. The Thematic Apperception Test (Morgan & Murray, 1935) has yet to meet all of these criteria, at least as it is generally used in current clinical practice. Even the 1916 Stanford-Binet, as well standardized as it was, relied on each examiner to assemble his or her own set of testing materials; that is,

Houghton Mifflin supplied the manual, test record forms, and a few printed materials, but numerous objects and supplies had to be assembled locally or ordered from the Stoelting Company. Standard test kits only became available for Forms L and M of the 1937 Stanford-Binet (Terman & Merrill, 1937).

Standard Materials and Procedures

All six of the aforementioned criteria are necessary, but obviously some are rather more central in generating the discriminative strength and power inherent in a well-constructed psychological test—the evidence of its established validity and the bases for detecting any particular invalid administration. Most of the other requirements permit some latitude. For example, in the original construction of the MMPI, the subject was presented with the test items on separate cards in a box and asked to sort them into categories of true, false, and cannot say. The test stimuli have gone through several changes in format (printed statements in a booklet or embossed in Braille, audiotaped and played aloud to the subjects, enacted in sign language on a videotape, or even scrolled on a computer display), as well as being translated into nearly 50 languages. However, research studies have shown that each of these alterations has preserved the standard meaning of the test stimuli. Similarly, several different response modalities have been introduced into the MMPI (marking answer sheets, pressing computer keys, and even oral responses recorded by an assistant); all permit immediate recording and avoid reliance on the memory of a test administrator. Studies of these modifications have not revealed any adverse effects on the motivational level of the subjects who are asked to take the inventory. Many of these alterations have actually served to increase the reliability of the recording and scoring of the responses obtained through different modes of administration. None of the modifications have substantially attenuated the test validities.

Optimal Motivation

A high level of motivation to perform the tasks in his intelligence scale was assured by Binet in his early scales by such a blatant technique as requiring the child to demonstrate his or her ability to remove the paper wrapper from a piece of candy (which the child could then eat). Terman, too, managed to keep interest high for his test subjects by means of attractive and intriguing toys and a variety of miniature objects that the subjects could manipulate. Motivation in completing the MMPI has been enhanced by the simplicity of the response format (Tversky, 1964), by the comfortable style of the wording of the items, and by the wide-ranging content of these statements (Fiske, 1969). More important, perhaps, is the fact that indicators have been built into the inventory that can signal any appreciable lack of interest or cooperation in completing the test appropriately. As important as optimal and appropriate motivation is to the successful execution of any test procedure, it is surprising to find that many published instruments still lack suitable methods of documenting a subject's test-taking compliance.

Immediate Recording

Today, with so many information-gathering procedures embedded in computer-orchestrated formats, it is perhaps puzzling to insist on the inclusion of immediate recording as one of the defining features of a test. Yet, there are still a large number of occasions when the "facts" of a person's history or behavior are filtered through the imprecise memory of an interviewer or observer. Ever-present errors of memory in recalling details of intake information recorded hours, days, or weeks after the interview are obviously the avenues by which errors of judgment are permitted to enter and to alter the permanent records on the individual.

Objective Scoring

Clear and unambiguous scoring standards are the hallmark of a psychological test. High interjudge reliability can be gained only by spelling out these criteria in sufficient detail to restrict drastically any room for idiosyncratic interpretation of the recorded behavior of the person taking the test. However, even with the responses of true or false clearly marked on an MMPI answer sheet and the direction of scoring unambiguously indicated by cut-outs on a scoring stencil, perfect reliability cannot be uniformly assured because of clerical errors arising from careless application of the stencils.

Appropriate Norms

The role of test norms in establishing the meaning of our measurement devices is more complex. Like the early efforts of the physicists to tie their measuring instruments to naturally occurring phenomena, we have desired our norming procedures to fulfill both measuring and interpretive functions. For example, international conventions were used by physicists to tie the length of the standard meter to "a fact of nature," one 10 millionth of the distance between the equator and the North Pole. A standard reference meter based on the known size of the earth at that time was marked out on a rod of platinum-iridium alloy and stored in a vault in Sèvres, France. Later, geophysical studies revealed that the actual distance was not 10 million meters but 10,002,288.3 meters. However, faced with this discordant information, physicists did not abandon the standard prototype meter (Asimov, 1960) but dropped all efforts to tie this unit to some "natural" constant. The meter still had innumerable uses without this excess meaning.

Similarly, when the MMPI was introduced outside of Minnesota, one of the arbitrary interpretive rules was the assumption that any score two standard deviations above the mean (i.e., a primed score or a value above a *T* score of 70) signified evidence of psychopathology. Subsequent research indicated that this cutting point produced an excess of false-positive interpretations and that these errors varied from one scale to another in the basic profile. At this point, it would have been possible simply to acknowledge these vagaries in the standard deviations developed by Hathaway and McKinley (1940) and propose that the rule of thumb involving the 70 *T*-

score cutting-point values be abandoned. Instead, inferences from the clinical and validity scores would be based on empirically determined cutting scores and clinical correlates. In fact, many computer-based interpretive systems for the MMPI had already made this kind of substitution. Nevertheless, in an effort to preserve the normality–abnormality implications inherent in MMPI *T* scores, the decision was made to establish entirely new norms for the traditional scales (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) and to introduce a set of statistical corrections to make uniform the extent of score deviations on each of the basic scales involved in coding patterning (Tellegen & Ben-Porath, 1992). However, some users of the MMPI have opted to retain the original standard deviation values for the profile scales, together with the rich interpretive lore that adheres to these older scaling standards and profile configurations.

Established Validity

It should be obvious that none of these modifications could have contributed any validity that the inventory did not originally garner from the steps taken by the authors in its construction. Here then is the key: empirical construction against powerful external criteria. There were very large differences between the subjects serving as the normative group for Hathaway and McKinley (1940) and those individuals grouped together on the basis of their common psychiatric conditions who served as criterion subjects. These differences reflected basic human characteristics that are vital for survival in society today. So were the differences that were manifested between the bright and dull pupils in the Paris school system who were studied by Binet (and those in the California schools studied by Terman). Those students differed in many characteristics essential for real-world success. Interestingly, the external criteria for these scale construction efforts were human judgments about humans; by the methods of item selection, small amounts of valid variance in the external criteria were captured incrementally into large concentrations of valid variance in the composite scales (Dahlstrom, 1985).

Often, it was not obvious to these pioneering psychologists just which tasks would prove discriminating against such criteria. Binet and his colleagues were unfortunately burdened by very poor conceptualizations about the nature of human cognitive processes. They had first to try out and reject reaction-time tasks, cephalic and anthropometric indexes, sensory thresholds (both absolute and discriminative), word associations, and imagery of various kinds. Only gradually did they come upon more suitable tasks for separating older from younger or brighter from duller pupils (in many instances from clues garnered by Binet in his systematic studies of the problem-solving methods used by his two daughters, Alice and Madeleine). Nevertheless, constant checking against observable, real-world differences in the subjects under scrutiny paid off handsomely in finding the particular tasks (and the appropriate ways of scoring them) that were needed.

Similarly, the authors of the MMPI began their work on the Minnesota Schedule with more than 1,000 test statements, gradually reducing to 550 the number of items that they would administer to their research subjects (Hathaway & McKinley, 1940). They believed this collection to be potentially useful in the various discriminations that they envisioned for the inventory. However, at first only 366 (or about two thirds) of the 550 proved useful for their purposes; perhaps equally useful ones were inadvertently discarded in their initial judgmental selection. These items in turn furnished the crucial separations required to construct workable scales that would stand up to changes in subjects, in circumstances and situations, in examiners, in languages, and even in cultures.

The robustness of the measures in the MMPI derives from the magnitude of the differences in emotional status manifested by the subjects used to generate the component scales. Previous instruments using scales assembled on the basis of clinical judgments about the component items failed to stand up to real-world criteria (Ellis, 1946, 1947). However, the differences captured in the valid variances of our empirically derived scales and instruments have turned out to reflect basic characteristics that are crucial to one's survival and successful functioning in society. Witness the findings in the long series of follow-up studies of the superior students identified by means of the Stanford-Binet in the elementary grades and by Terman (1925) in the high schools of California. Innumerable indexes of the later successes achieved by these youngsters have been documented in science, medicine, and the arts (Terman, 1959). Since Terman's death, these investigations have been continuing and are currently under the direction of L. J. Cronbach.

One interesting and rather unexpected consequence of serving as a subject in such an intensive and long-term investigation is illustrated by the reports back in the 1950s from students of Professor Irving Lorge in his class on psychometric methods at Teachers College, Columbia University. They indicated that to drive home his point on how hard students would have to work on the assignments in this course, he would remind each class that he was one of "Terman's geniuses." His career amply documented the fulfillment of this prognostication from the S-B: early high-school graduation, PhD at 25 from Columbia, full professor at 33, and probably best known for his collaboration with Robert Thorndike on the Lorge-Thorndike dictionary. How many other members of this cohort were stimulated into special accomplishments by the knowledge that they had been singled out early, through nomination by a teacher and screening by a psychological test, as a person who would soon achieve eminence? How many others built an important aspect of their self-image around this kind of prestigious label?

There are many other examples of the broad implications of psychological test results. In the early 1960s, Earl Baughman and I (Baughman & Dahlstrom, 1968) gathered extensive test data on all of the pupils in a rural school system near Chapel Hill, North Carolina, using among other instruments the Stanford-Binet (Terman &

Merrill, 1960), the Stanford Achievement Tests (SAT; Kelley, Madden, Gardner, Terman, & Ruch, 1953), and for the eighth-grade students a tape-recorded administration of the MMPI (Butcher & Dahlstrom, 1964). With our test findings we were able to document for the superintendent of the school system his need for extra funds from the state to provide for special education classes for many of these students. We were also able to identify, for the first time in this community, students who were eligible for the Governor's School for the Academically and Artistically Talented.

These test data were later used by Lowman, Galinsky, and Gray-Little (1980) to carry out a 10-year follow-up study of the former eighth-grade students in our project. In their study they examined the relationships of Stanford-Binet IQs, Stanford Achievement Test scores, and MMPI Disturbance Index (DSI) values (Cooke, 1967) to various indicators of success in the subsequent careers of these boys and girls over the intervening decade: how far they went in school, how much money they were earning, what level of occupation they had reached, whether they were married and how many children they had, and whether they had remained in the home community or migrated to urban settings or out of the state altogether. Intelligence quotients from the 1960 version of the S-B forecast to a highly significant degree their final level of schooling, the level of occupation that they achieved, and even the level of income they were earning a decade later. The scores on the SAT were not as closely related to these outcomes as the S-B IQ values, except for predictions about Black male students, for whom the SAT scores turned out to be somewhat better predictors than did the S-B IQ values. The level of pathology reflected in Cooke's DSI values from the MMPI demonstrated only marginal significance against these outcomes but served to strengthen various predictions in multiple regression analyses.

As another example, Hathaway and Monachesi (1953) documented the increase in accuracy provided by the MMPI over base rates of juvenile delinquency in students enrolled in the schools of Minneapolis. Although roughly one in five boys (22%) showed up in police records between the time they were tested in 9th grade until the follow-up ended in 12th grade, more than 40% of boys with "exciter" scales (Scales 4, 8, and 9) elevated in their test profiles had come to the attention of the police. (A similar doubling of the base rates was also reported for the girls with exciter scale elevations, from 11% to more than 22%.) Even more surprising was the information obtained on further follow-up by Briggs, Wirt, and Johnson (1961). If indicators of family disorganization were added to the MMPI data, nearly 9 out of 10 (88%) of the boys with a 489 configuration on the MMPI, who lived in disrupted families, could have been identified in the 9th grade as later juvenile delinquents. Interestingly, the data in the 9th-grade testing survey did not predict very well those men with delinquent histories who continued on into adult criminal careers, in contrast to the majority of these men who soon matured into solid citizens with

no further involvement with the police. This latter difference appeared to be more closely related to the various experiences they encountered at the hands of the authorities in the juvenile justice system.

Gottesman and his colleagues (Hanson, Gottesman, & Heston, 1990) have also gathered data on the subjects in these same ninth-grade cohorts when they subsequently appeared in mental health facilities around the state of Minnesota. Through these studies (Moldin, Gottesman, Rice, & Erlenmeyer-Kimling, 1991), they have established the predictive power of additional MMPI configurations in the test records obtained when the subjects were about 14 years old. Similarly, Hoffman, Loper, and Kammeier (1974) demonstrated the relationship between MMPI patterns in the test results of entering freshmen at the University of Minnesota and the strong likelihood that students with these test score configurations would later enter state rehabilitation programs for substance abuse.

Using data gathered on groups of medical school and law school students from a quarter of a century earlier, investigators at the Behavioral Medicine Research Center at Duke University (Barefoot, Dahlstrom, & Williams, 1983; Barefoot, Dodge, Peterson, Dahlstrom, & Williams, 1989) have shown a strong relationship between Cook and Medley (1955) Hostility scale scores and increased morbidity and mortality from coronary artery disease in these physicians and lawyers. Behaviors as specific as the extent of cigarette smoking (and difficulties in smoking cessation) or the extent of regular physical exercise can also be predicted from either the standard clinical scales or special research scales in the MMPI (Siegler et al., 1991, in press). Test-based evidence of problems in health and survival that may be encountered decades later should be put to constructive use in guiding appropriate intervention efforts to alter adverse interpersonal styles or pathogenic behavior patterns. In this context it is fitting to remind ourselves of the warning that J. Robert Oppenheimer gave psychologists back in the mid-1950s:

In the last ten years the physicists have been extraordinarily noisy about the immense powers which, largely through their efforts, but through other efforts as well, have come into the possession of man, powers notably and strikingly for very large-scale and dreadful destruction. We have spoken of our responsibilities and of our obligations to society in terms that sound to me very provincial, because the psychologist can hardly do anything without realizing that for him the acquisition of knowledge opens up the most terrifying prospects of controlling what people do and how they think and how they behave and how they feel. This is true for all of you who are engaged in practice, and as the corpus of psychology gains in certitude and subtlety and skill, I can see that the physicist's pleas that what he discovers be used with humanity and be used wisely will seem rather trivial compared to those pleas which you will have to make and for which you will have to be responsible. (Oppenheimer, 1956, p. 128)

Summary

The samples of behavior that psychologists collect in the brief time that an hourglass takes to empty have been shown to reveal basic aspects of ability, personality, and

temperament that are operative over long spans of an individual's life. Proper gathering of these data by means of well-executed administrations of standardized test instruments can provide gatekeepers with invaluable information to minimize risks of errors of judgment in decisions about their clients and increase the range of predictions that can have large consequences in the lives of those with whom they deal. Many of these predictions can be used to guide and focus appropriate interventions designed to minimize the risks of later psychopathological or psychosomatic disorders, substance abuse or criminal involvement, or even premature deaths. Gatekeepers of the future can be armed with a whole new array of screening and prognostic indicators to help improve the lives of their clients. It should be equally clear that information based on poorly constructed and inadequately standardized tests (or poorly executed administrations of even our best instruments) can serve to mislead and distort clinical judgments based on biased samples of behavior. Continuing efforts must be devoted to the enhancement of the validities of our psychological tests and constant vigilance maintained against misuses or abuses of these powerful devices. The consequences, constructive or destructive, are too large to neglect.

REFERENCES

- Asimov, I. (1960). *Realm of measure*. Boston: Houghton Mifflin.
- Barefoot, J. C., Dahlstrom, W. G., & Williams, R. B. (1983). Hostility, CHD incidence, and total mortality: A twenty-five year follow-up study of 255 physicians. *Psychosomatic Medicine*, 45, 59-63.
- Barefoot, J. C., Dodge, K. A., Peterson, B. L., Dahlstrom, W. G., & Williams, R. B. (1989). The Cook-Medley Hostility Scale: Item content and ability to predict survival. *Psychosomatic Medicine*, 51, 46-57.
- Baughman, E. E., & Dahlstrom, W. G. (1968). *Negro and White children: A psychological study in the rural South*. San Diego, CA: Academic Press.
- Binet, A., & Simon, T. (1907). *Les enfants anormaux*. Paris: Armond Colin.
- Briggs, P. F., Wirt, R. D., & Johnson, R. (1961). An application of prediction tables to the study of delinquency. *Journal of Consulting Psychology*, 25, 46-50.
- Butcher, J. N., & Dahlstrom, W. G. (1964). *Comparability of the taped and booklet versions of the MMPI*. Unpublished manuscript, University of North Carolina at Chapel Hill.
- Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Manual for administration and scoring the MMPI-2*. Minneapolis: University of Minnesota Press.
- Cook, W. W., & Medley, D. M. (1955). Proposed hostility and Pharisaiic-virtue scales for the MMPI. *Journal of Applied Psychology*, 38, 414-418.
- Cooke, J. K. (1967). MMPI in actuarial diagnosis of psychological disturbance among college males. *Journal of Counseling Psychology*, 14, 474-477.
- Dahlstrom, W. G. (1985). The development of psychological testing. In G. A. Kimble, & K. Schlesinger (Eds.), *Topics in the history of psychology* (Vol. 2, pp. 63-113). Hillsdale, NJ: Erlbaum.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243, 1668-1674.
- Ellis, A. (1946). The validity of personality questionnaires. *Psychological Bulletin*, 43, 385-440.
- Ellis, A. (1947). Personality questionnaires. *Review of Educational Research*, 17, 53-63.
- Exner, J. E. (1974). *The Rorschach: A comprehensive system* (Vol. 1). New York: Wiley.
- Exner, J. E. (1978). *The Rorschach: A comprehensive system* (Vol. 2). New York: Wiley.
- Eyde, L. D., Kowal, D. M., & Fishburne, F. J. (1991). The validity of computer-based test interpretations of the MMPI. In T. B. Gutkin & S. L. Wise (Eds.), *Buros-Nebraska Symposium on Measurement and Testing: The computer and the decision-making process* (Vol. 4, pp. 75-124). Hillsdale, NJ: Erlbaum.
- Fiske, D. W. (1969). Subject reactions to inventory format and content. *Proceedings of the 77th Annual Convention of the American Psychological Association*, 4, 137-138.
- Hanson, D. R., Gottesman, I. I., & Heston, L. L. (1990). Long-range schizophrenia forecasting: Many a slip twixt cup and lip. In J. Rolf, A. S. Masten, D. Cicchetti, K. H. Nuechterlein, & S. Weintraub (Eds.), *Risk and protective factors in the development of psychopathology* (pp. 424-444). Cambridge, England: Cambridge University Press.
- Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): 1. Construction of the schedule. *Journal of Psychology*, 10, 249-254.
- Hathaway, S. R., & Monachesi, E. D. (1953). *Analyzing and predicting juvenile delinquency with the MMPI*. Minneapolis: University of Minnesota Press.
- Hoffman, H., Loper, R. G., & Kammeier, M. L. (1974). Identifying future alcoholics with MMPI alcoholism scales. *Quarterly Journal of Studies of Alcohol*, 35, 490-498.
- Kelley, T. L., Madden, R., Gardner, E. F., Terman, L. M., & Ruch, G. M. (1953). *Stanford Achievement Test: Manual*. Yonkers-on-Hudson, NY: World Book.
- Lowman, J., Galinsky, M. D., & Gray-Little, B. (1980). *Predicting achievement: A ten-year followup of Black and White adolescents*. Chapel Hill, NC: Institute for Research in Social Science.
- Moldin, S. O., Gottesman, I. I., Rice, J. P., & Erlenmeyer-Kimling, L. (1991). Replicated psychometric correlates of schizophrenia. *American Journal of Psychiatry*, 148, 762-767.
- Morgan, C. C., & Murray, H. A. (1935). A method for investigating fantasies: The Thematic Apperception Test. *Archives of Neurology and Psychiatry*, 34, 289-306.
- Oppenheimer, R. (1956). Analogy in science. *American Psychologist*, 11, 127-135.
- Rorschach, H. (1921). *Psychodiagnostik*. Berne, Switzerland: Birken.
- Siegler, I. C., Blumenthal, J. A., Costa, P. T., Dahlstrom, W. G., Peterson, B. L., Barefoot, J. C., & Williams, R. B. (1991). Personality prediction of exercise in the UNC Alumni Heart Study. *Psychosomatic Medicine*, 53, 220-221.
- Siegler, I. C., Peterson, B. L., Barefoot, J. C., Harvin, S. H., Dahlstrom, W. G., Kaplan, B. H., Costa, P. T., & Williams, R. B. (in press). Using college alumni populations in epidemiologic research: The UNC Alumni Heart Study. *Journal of Clinical Epidemiology*.
- Tallent, N. (1992). *The practice of psychological assessment*. Englewood Cliffs, NJ: Prentice Hall.
- Tellegen, A., & Ben-Porath, Y. S. (1992). The new uniform T scores for the MMPI-2: Rationale, derivation, and approach. *Psychological Assessment*, 4, 145-155.
- Terman, L. M. (1916). *The measurement of intelligence*. Boston: Houghton Mifflin.
- Terman, L. M. (1918). Expert testimony in the case of Alberto Flores. *The Journal of Delinquency*, 3, 145-164.
- Terman, L. M. (Ed.). (1925). *Genetic studies of genius: Vol. 1. Mental and physical traits of a thousand gifted children*. Stanford, CA: Stanford University Press.
- Terman, L. M. (Ed.). (1959). *Genetic studies of genius: Vol. 5. The gifted group at mid-life*. Stanford, CA: Stanford University Press.
- Terman, L. M., & Merrill, M. A. (1937). *Measuring intelligence*. Boston: Houghton Mifflin.
- Terman, L. M., & Merrill, M. A. (1960). *Stanford-Binet Intelligence Scale: Manual for the third revision* (Form L-M). Boston: Houghton Mifflin.
- Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology*, 1, 386-391.
- Wechsler, D. (1939). *The measurement of adult intelligence*. Baltimore: Williams & Wilkens.
- Whitty, P. A., & Jenkins, M. D. (1935). The case of B---: A gifted Negro child. *Journal of Social Psychology*, 6, 117-124.
- Yerkes, R. M., Bridges, J. W., & Hardwick, R. S. (1915). *A point scale for measuring mental ability*. Baltimore: Warwick & York.