ASSESSING THE CANDIDATE AS A WHOLE: A HISTORICAL AND CRITICAL ANALYSIS OF INDIVIDUAL PSYCHOLOGICAL ASSESSMENT FOR PERSONNEL DECISION MAKING

SCOTT HIGHHOUSE
Department of Psychology
Bowling Green State University

Although individual assessment is a thriving area of professional practice in industry, it receives little, if any, attention from textbooks on industrial psychology or personnel management. This article is an attempt to establish individual assessment's place in the history of personnel selection, and to examine why the practice has survived despite receiving little attention in research and graduate training. It is argued that the clinical, holistic approach that has characterized individual-assessment practice has survived primarily because the "elementalist" testing approach, focusing on traits and abilities, has often been dismissed as inadequate for addressing the complexities of the executive profile. Moreover, public displeasure with standard paper-and-pencil testing in the 1960s and 1970s made the holistic approach to assessment an attractive alternative. The article contrasts individual assessment practice with the current state of knowledge on psychological assessment and personnel decision making. Like psychotherapy in the 1950s, individual psychological assessment appears to have achieved the status of functional autonomy within psychology.

The label individual assessment in personnel decision making has come to define a set of psychological procedures used by trained psychologists to make inferences about a person's fit with a job (Hansen, 1991; Jeanneret & Silzer, 1998). Although practices vary widely from assessor to assessor, it is perhaps best to conceptualize individual assessment as a loosely defined set of procedures, similar to practices developed and performed in clinical and counseling settings, used to make recommendations for higher level hires. The practice is commonly used for evaluating candidates for executive positions or assessing suitability for

I am grateful to Maggie Bro: ks-Laber, Robert Guion, Charlie Reeve, Mike Zickar, and three anonymous reviewers for their valuable contributions to the final product. This work was supported by a financial grant from the Patricia and Olin Smith Faculty Development Fund at Bowling Green State University.

Correspondence and requests for reprints should be addressed to Scott Highhouse, Department of Psychology, Bowling Green State University, Bowling Green, OH 43403; shighho@bgnet.bgsu.edu.

COPYRIGHT © 2002 PERSONNEL PSYCHOLOGY, INC.
specialized assignments (e.g., law enforcement agent). Characteristic of these positions is a situation where performance is difficult to define and relatively few people occupy the roles. Thus, success criteria are often unavailable. The difficulty in determining assessor success, along with the high premiums often paid to assessors, has made this an inviting area of practice for consulting psychologists and charlatans alike. It has also made individual assessment an area of inquiry that academics have shied away from (cf. Ryan & Sackett, 1989). Nevertheless, this type of selection is an area that constitutes a sizable slice of the professional practice pie in personnel psychology (Jeanneret & Silzer, 1998). As Guion (1998) noted, “individual assessment is alive and well as an area of professional practice, if not as an area of research” (p. 631).

Individual assessment is distinguished from traditional personnel selection in that it typically relies heavily upon psychological measures and unstructured interviews, and data are often integrated via clinical intuition (Ryan & Sackett, 1998). As such, it is often viewed as more an art than a science, and the focus of assessment is on the candidate as a whole, rather than on individual abilities or traits. The notion of assessing the total person has a storied tradition in 20th century applied psychology. This article documents the development and survival of this holistic approach to selection, and examines its underlying assumptions and associated practices. The article begins with an historical treatment of the work conducted by psychologists and psychiatrists in Germany, England, and the United States during World War II. Particular attention is paid to the work conducted by Henry Murray and his colleagues for the Office of Strategic Services. Although this work is commonly cited as seminal in the development of the assessment-center method, I suggest that Murray’s ideas provided the intellectual foundation for individual psychological assessment in organizations. An attempt is made to trace Murray’s influence after the war, and to explore how the holistic approach has survived in spite of neglect from the personnel testing community. Finally, the “rediscovery” of individual assessment in the 1990s is documented, and its practice is critiqued in light of current research findings in psychological science.

Identifying German Character

The earliest application of the total-person concept of assessment for selection occurred in the German Army during World War II. In response to the poor showing of the German Army in World War I, German High Command established a program for assessment and selection of officers and specialists. In 1936, there were 15 German Army psychological laboratories—84 military psychologists processed more than
40,000 candidates a year. By 1941, there were approximately 500 psychologists working for the wehrmacht, or armed forces (Fitts, 1946). Psychological work in the army was under the direction of Max Simoneit, head of the center for psychological research at the University of Berlin. Simoneit, regarded as the “dean” of German military psychology (Ansbacher, 1941a), advocated ausdrucksanalyse (i.e., analysis of expression) as the best means of identifying a person’s character (Simoneit, 1954). Expression analysis and characterology dominated German psychology during the Nazi period (Geuter, 1987), and was aimed at eliminating “atomistic” distinctions between abilities and traits (Wiggins, 1973). The global concept of “practical disposition” replaced the earlier German concept of “pure intellect” and the American concept of multiple abilities measured by standardized tests (see Guilford, 1940). Indeed, psychometric methods, such as tests of ability or reaction time, were stereotyped as “Jewish” (Geuter, 1987).

Simoneit was highly influenced by the work of Ludwig Klages and Philipp Lersch, conducted in the 1920s and 1930s. Klages drew a distinction between the coping activity of the geist (mind) and the expressive activity of the seele (soul). A special interest of Klages was handwriting, which he believed reflected a struggle between expressive and coping behavior. Whereas, when writing, a person is deliberately conveying his thoughts via paper and pen (i.e., coping), the person is at the same time seen by Klages as betraying his energy, aggressiveness, hostility, fear, ambition, and rigidity (Allport, 1961). Lersch believed that the structure of character was based on a model consisting of feelings directed by a superstructure of will and thought (Geuter, 1987). The work of Lersch focused on facial expressions and their possible meanings. His emphasis on the importance of facial expressions clearly influenced many of the techniques developed by Simoneit.

Simoneit believed that the candidate’s manner of performing a task was more important than his actual achievement on the task (Ansbacher, 1941b). That is, the composure with which someone approached a task and overcame obstacles was regarded as more diagnostic of the person’s character than the performance on the task itself. Simoneit’s assessment of officer candidate character was composed of four parts: (a) intelligence analysis, (b) action analysis, (c) expressive analysis, and (d) life history interview (Simoneit, 1938). These are discussed in turn.

Practical intelligence was valued over what was measured by psychometric intelligence tests. For example, practical judgment was assessed

---

1 Fitts (1946) speculated that one reason for the lack of interest in objective testing lay in the belief that there were unique (read superior) qualities of the German mind and character that were immune to analysis via standard tests.
by examining essays written in response to situations such as the one described by Ansbacher (1941a, p. 381):

Your sports club is planning a trip to Berlin to attend the Germany–England football game, which will take place in 2 weeks. You have been entrusted with the preparations and entire management of the trip. What do you intend to do? Describe your plan in detail.

Thus, we see what is probably one of the first uses of situational-judgment testing in an employment setting. In scoring such items, no single action was considered correct, but the reasons for each action were studied for indications of planning ability and good judgment (Fitts, 1946). One example of Simoneit's action analysis was to have the candidate respond in a certain way to specific stimuli, to pull a lever with the right hand when a red light is shown, and with the left hand when a blue light is shown. Action analysis techniques might involve the use of both hands and feet, responding to both light and sound. The procedures were concerned with identifying both coordination and tolerance for stress. Stress tolerance might also be assessed by having the candidate walk on a free swinging pole and, at the same time, throw a rope over several hooks on his left and right—all the while being criticized and ridiculed severely by the assessors.

The analysis of expression was perhaps the most unusual, and was based on measurement of facial expressions, body language, verbal emissions, and handwriting. In one of Simoneit's tasks, the candidate was required to pull an expanding-spring exerciser that emitted increasing amounts of electrical current as it was pulled. The facial reactions of the candidate were recorded by a hidden camera, and a team of assessors observed his general behavior. Other character tests developed principally by Lersch included the analysis of voice and facial expression as the candidate was physically exhausted by an obstacle course impossible to complete (Fitts, 1946), and the observation of candidates while they groomed in front of a mirror (Geuter, 1987). Also included in the analysis were leaderless group discussions, and an “orders test” in which candidates were required to direct assistants in the accomplishment of a task. After the testing was completed, a final interview was conducted to draw a final impression of the “total personality,” taking into account impressions developed during the testing sessions. The interview was unstructured and explored attitudes and biographical items. Special attention was given to cultural and socioeconomic background, and to knowledge of literature, art, and philosophy (Fitts, 1946).

The final evaluation of the candidate was made on the basis of intuition. Examiners were often encouraged to vary testing procedures and to give special weight to their preferred tests. According to Fitts (1946),
there was little concern over differences between examiners, and little appreciation for the concepts of reliability and standardization. Deductions were based on preferred theories of personality structure, along with the matching of candidate profiles with profiles of military heroes developed from case studies. Despite the divergence of these procedures from accepted practice in American psychology, the evidence suggests that the German procedures were highly publicized and viewed as innovative by psychologists in both England and the United States (e.g., Pratt, 1941).

_Anglo-Analytic Assessment_

By 1941, the British Army set up the War Office Selection Board (WOSB, pronounced "was be") in response to the shortcomings of the old interview boards being used to identify officers. Prior to instituting the WOSBs, failure rates at Officer Cadet Training Units were at 20% to 30% (H. Murray, 1990), and many recruits who might have succeeded in training were rejected by the interview boards on the basis of such factors as which grammar school they attended or their perceived "socialist opinions" (Vernon & Parry, 1949). The WOSBs were staffed by a president, a psychiatrist, military testing officers, and a psychologist. Whereas the psychiatrists conducted interviews to make character assessments, and the officers conducted the situational tests to make leadership judgments, the psychologists were confined initially to administering intelligence and projective tests. The core group of psychiatrists and psychologists responsible for developing the WOSBs came primarily from the Tavistock Institute of Medical Psychology, and were later identified as the "Tavistock Group."² The practices of the Tavistock psychologists were more aligned with the practices of their psychiatrist colleagues than with the "headquarters" psychologists in the War Office who were trained in industrial psychology and psychometry (Vernon & Parry, 1949).

The approach of the Tavistock Group is difficult to classify, but might be thought of as a type of social psychiatry. Members were influenced by the emergent object-relations approach, a British version of psychoanalysis, which emphasized relationships rather than instinctual drives and psychic energy (see Sutherland, 1963). Also influential were Kurt Lewin's field theory, and J. L. Moreno's work in psychodrama. Lewin's view of personality organization stressed that, in order to understand any given piece of behavior, one must look at the larger system of which it

---

²The Tavistock Clinic was formed after the first world war as an outpatient clinic for dealing with war-related neuroses. Following WWII, the Tavistock Institute of Human Relations was formed to combine psychoanalysis with psychosocial studies.
PERSONNEL PSYCHOLOGY

is a part (see DeRivera, 1976). In the same spirit, Moreno had done extensive experimentation with "sociometric" assessments as alternatives to traditional psychometric approaches. The sociometric approach emphasized the assessment of people in real-life situations. Subjects were asked to act out, in front of an audience, roles and situations as they occur in real life. According to Moreno, "Many traits that indicate personality difficulty are disclosed: anxieties, stage fright, stuttering, fantasies, unreasonable attitudes, and so on" (1941, p. 386). Moreno's ideas fit well with the WOSB goal of assessing the whole man in his natural social environment.

Interestingly, the role-playing tests, for which the WOSBs had become synonymous, were originally introduced by the military higher ups as a "cover" for the work of the psychological technicians (Vernon & Parry, 1949). There was much initial skepticism of having officers selected by doctors who were obsessed with ferreting out sexual complexes. Therefore, the military testing officer (MTO) was introduced to apply tests of a military nature, which were expected to impress the army while the real work of analyzing officer character was being conducted by the psychiatrists and psychologists. From the situational tests, however, evolved the leaderless group tests. The leaderless group method, in which a group was left to its own devices in coping with situations presented by the MTO, was conceived by W. R. Bion (1946). Bion and his colleagues did not accept the notion of traits as constant qualities of a person that exist independent of the context. Instead, they believed that personality, as an organized whole, must be assessed in real-world situations where the total configuration of traits was needed for success.

Unlike the leaderless group discussions used by the Germans, the leaderless group tests were action oriented. For example, a group of six to eight candidates might be instructed to build a makeshift bridge for crossing a body of water, or to move a heavy object over a series of obstacles. Bion saw these activities as producing in the candidate a tension between maintaining personal relationships, while simultaneously considering his own interests (Bion, 1946). Some candidates would become mere "passengers," giving up their own desire to shine, and others would subordinate the groups' interests to their own (Vernon & Parry, 1949). In reflecting on his experiences as a (failed) candidate of a WOSB, Edmund Mallett recalled:

We were in groups of about six when doing the outdoor tests. One was an area that was supposed to be contaminated and we had to get across this using rope, poles, etc. We had to work out the method of assembling these while officers watched us, taking notes. There were the born leaders who attempted to tell others what to do; there were those who more quietly thought out the problem; and there were those who said or did little. I
Although most candidates thought that they were being evaluated on the basis of problem accomplishment or forcefulness, Bion was most interested in the quality of the individual's social relationships. As the candidate learned that he could only demonstrate his abilities through the medium of others, it was expected that the capacity for mature, independent social relationships would emerge (Sutherland & Fitzpatrick, 1945).

Over time, the use of different assessors observing different aspects of candidate behavior (i.e., the psychiatrist in the interview, the MTO in the leaderless tests, and the psychologist in the standard testing situation) violated the sensibilities of the assessors who valued the rounded judgment based on observation of the total person (Sutherland & Fitzpatrick, 1945). The solution to this problem was to combine the three observers into an observer team. This team would share observations and background data on the candidates. In addition, they would all observe behavior in the leaderless group tests and regularly convene to share impressions. William J. Morgan, an American who served as an assessor in the WOSB at Pemberley in England, noted that, "We were constantly modifying the tests in use with the purpose of making them less structured" (1957, p. 68). According to Morgan, candidates were seen as more likely to reveal their true character when they had more choice in solving a problem. These modifications to the situational tests were clearly inconsistent with the psychometric school, which emphasized standardization and independence of observations between assessors and across situations. The modifications were consistent, however, with the ideas being developed at the same time by Henry A. Murray in the United States.

**Cloak and Dagger Types**

A milestone in individual assessment was the development in the United States of a program to select spies and saboteurs for the Office of Strategic Services (OSS) during World War II. Under the direction of General William J. Donovan, the OSS needed to identify men and women to serve as agents to, among other things, conduct destructive operations behind enemy lines, train resistance groups, and "disintegrate" the morale of enemy troops (OSS, 1948). In addition, the people selected to serve overseas needed to have the tolerance for stress to hold up under physical torture by the Gestapo. The psychologists chosen to develop the assessment program included such notable figures as Henry
Murray, Donald Fiske, Edward Tolman, Eugenia Hanfmann, and Donald MacKinnon. The orientation of these pioneers was primarily personality theory and psychoanalysis. Conspicuously absent from the group were personnel psychologists.

Although the assessment program was a collaborative effort, the influence of Henry Murray was significant (MacKinnon, 1977). Murray, by his own admission, had no formal education in psychology (H.A. Murray, 1967). Murray's training was in medicine, but he became preoccupied with understanding the inner workings of the mind and was strongly influenced by the work of Carl Jung. Murray had become well known for his development of the Thematic Apperception Test (TAT), a projective device for revealing unconscious motivations, along with the development of a taxonomy of human needs (Murray, 1938). He was considered a maverick who rejected psychology's focus on narrowly restricted forms of behavior and its preoccupation with individual traits. Murray's approach to personality assessment involved applying holistic principles by inferring general tendencies, traits, and their interrelations from a number of specific signs. This differed from the psychometric model that involved combining and weighting objectively measured isolated traits. Thus, the holistic approach called for a subjective diagnosis of personality (Murray & MacKinnon, 1946).

The holistic approach favored by Murray, and adopted by the OSS assessment staff, relied heavily upon the ability of the psychologist to observe patterns in candidate behaviors and to integrate those observations into a judgment of effectiveness. As the OSS report stated: "And so it could be said that this method calls for the improvement of the psychologist as observer and interpreter rather than the improvement of mechanical instruments and test materials" (OSS, 1948, p. 40). The staff argued that one's score on a test is no more a representation of effectiveness in the field than the measurement of one bodily movement is a representation of a proper golf swing. The holistic method of assessing total integrated processes, rather than partial isolated processes, was seen as appropriate to the OSS case because the staff was faced with the problem of discriminating between candidates who fell in the upper half of the distribution of general competence. Because candidates were selected on the basis of demonstrated skill, OSS standards were higher than those of the majority of the military service. Thus, tests that had previously been shown to distinguish people representing the entire range of ability were not necessarily seen as useful for the OSS program (Murray & MacKinnon, 1946).

Another reason for the rejection of the traditional model of assessment was the perceived necessity of judging social relations among the candidates. The OSS psychologists and psychiatrists were expected to
assess a candidate's ability to get along well with others, organize the activities of others, and evoke respect. No existing standardized measures were seen as adequate in addressing these issues (OSS, 1948). The requirements on which all candidates were assessed included: (a) motivation for assignment, (b) energy and zest, (c) practical intelligence, (d) emotional stability, (e) social relations, (f) leadership ability, and (g) security (i.e., secrecy). Although ratings were made on each of these dimensions, the OSS staff was clear on the point that the “whole personality and its parts are mutually dependent” (p. 43). Whereas the dimensions could outline the attributes needed in a candidate, only the psychologist could determine how they were combined in any one person. Thus, achieving a final, overall assessment of the candidate was the primary goal of the assessment program.

The first assessments took place at Station S on a secluded farm outside of Washington, D.C. (A shorter version of Station S, known as Station W, was later implemented for handling the overflow of candidates). Candidates underwent more than 3 days of assessments that included projective tests, interviews, situational tests, paper-and-pencil examinations, psychodrama, and casual observation. Upon entering Station S, the candidates were stripped of their personal belongings, dressed in army fatigues, and instructed to create false identities that they would maintain throughout the duration of the program. Candidates were only allowed to reveal their true identities during “X” conditions, where they met individually with psychologists to discuss their life histories. The reason for this was not only to determine the candidate’s ability to keep secrets, but also to protect candidates from infiltrators into the program who might use personal information against them in the field (MacKinnon, 1977). Although a complete description of the procedures used at Station S is beyond the scope of this manuscript (see OSS, 1948, for a comprehensive treatment of the assessment program), it is worthwhile to focus on some of the procedures used so the reader may appreciate the execution of the clinical, holistic approach advocated by Murray.

The primary purpose of the projective tests was to provide hints about a candidate’s personality that, together with the candidate’s written personal history, could be followed up in the personal interviews. Although the OSS staff was cautious about overinterpreting responses to the projectives, they clearly believed that the tests were capable of revealing aspects of the unconscious that would not be revealed otherwise. For example, the projective questionnaire included 12 items that called for a narrative response. One item, “What kinds of things do you most dislike to see people do?” was based on the theory that the things a person most dislikes about another are “the very ones that he has tried to suppress in himself” (OSS, 1948, p. 90). Other items were aimed at
uncovering feelings of inadequacy, loneliness, and frustration. The most popular of the projective devices was a 100-item sentence completion test. Items referred either to the candidate (e.g., “I admire...”) or to another person (e.g., “Charlie was happiest when...”). The latter approach was based on the assumption underlying all projective procedures that, in describing others, one reveals his or her own motivations. The Sentence Completion Test was not scored, but interpreted without standardized procedures. Attention was primarily given to repetitive responses and to responses that were rare when compared to the responses of other candidates. According to the OSS report, the prestige of the Sentence Completion Test grew steadily as the interpreters gained experience with it. Indeed, it was the only projective test that remained in use throughout the history of the OSS program.

The use of situational tests was explicitly borrowed from the English WOSBs. Also borrowed was the idea of planting “stooges” in the situational tests in order to increase the stress level and test leadership qualities. One such test was the Construction Test, which involved having the candidate instruct two “privates” in how to build a miniature house frame out of a pile of large poles. The candidate was given 5 minutes to lead the men while the privates feigned ignorance or obstinacy. Howard Hunt, of Watergate fame, recalled his experience in this exercise as an OSS candidate:

When my turn came, I entered the barn and found the two privates sitting dejectedly on the floor, not even looking up when I came in. I took a look at the poles and told the men to lay certain of them in a rectangle and to raise uprights at the corners. With maddening slowness the privates got to their feet and shuffled about their work... The uprights finally were raised when one of the men stumbled against the frame and down it came. This they regarded as a great joke... One of them muttered that he hated all officers (Hunt, 1974, p. 34).

Although this test often evoked strong emotions from candidates, sometimes even leading to fist fights with the privates (MacDonald, 1947), the assessors were never clear on what behavior from the candidates should be regarded as appropriate. The OSS report confessed, “Just what was measured in this situation is difficult to state exactly” (OSS, 1948, p. 111).

Another situational assessment, which was considered among the most revealing, was the Improvisation. Based on Moreno’s psychodrama techniques, the Improvisation was a dramatic situation chosen specifically for each candidate to act out. No two situations were the same, and they were invented by the assessment staff to prove or disprove “critical hunches” they had developed about the candidate’s personality. The situations might be designed to assess, for example, a candidate’s tact in
dealing with superiors, proneness to guilt feelings, reactions to accusations of dishonesty, or attitudes toward minority groups. According to the OSS report: "the men were revealing a good deal about their modes of conducting interpersonal relationships. The test was then developed along much broader lines than had been originally intended" (OSS, 1948, p. 170). A brief experiment with rating personality variables revealed in the improvisations was abandoned entirely by the assessors, as they believed that simply listening to each plot as it unfolded led to greater insight (OSS, 1948).

A description of the OSS assessment program would not be complete without mentioning the Interrogation Test. Each candidate was brutally interrogated by several experienced questioners whose aim it was to unravel a hastily invented false identity. The purpose of the test was to assess the candidate's ability to tolerate severe emotional strain. According to the OSS report, the candidate was not allowed to relax:

He was made to sit upright in a hard chair. If he was smoking, he was told to stop. If he crossed his legs, he was told to uncross them. If he lowered his head to avoid the light, he was commanded to look up. If he wore glasses, he was directed to take them off (OSS, 1948, p. 136).

After 10 minutes of interrogation, the candidate was informed, "you have failed the test." The candidate was then told to report to a staff member in another room. This second interview was designed to put the candidate at ease and encourage discussion of reactions to the interrogation. The hidden purpose of the second interview was to entice the candidate to break cover. Many candidates were said to have suffered from crippling anxiety attacks and, thus, the aim of the procedure was not only to assess stress tolerance but also to assess candidate ability and motivation for assignment to the OSS.

The overall evaluation of candidates was done via continuous discussion among the assessment staff who ultimately would arrive at a consensus judgment—without resorting to statistically averaging the dimension ratings. Indeed, the report is clear in admitting that the dimension ratings were not dependable indicators of effectiveness in specific situations (OSS, 1948, p. 205). This is because scores in situations scheduled late in the program were consciously influenced by the cumulative evaluation of the candidate. The goal was to make all information acquired about a candidate available to each assessor at all times. This was seen as the logical outcome of their attempt to arrive at a comprehensive personality picture. Furthermore, considerable weight was given to inferences made about candidate personalities on the basis of casual interactions between assessors and candidates as they ate meals together, relaxed, or played softball. Also influential was the "never mentioned" and "never
named" liquor test. During some of the formal testing procedures (i.e., improvisations and debates), candidates would be offered hard liquor, and care was taken to ensure that their glasses were kept full. The observation of candidates under the influence of alcohol was seen as helpful in revealing new aspects of personality that could either round out earlier impressions or modify them completely.  

_Reaction of the Elementists_

Although the OSS report levied considerable criticism toward traditional approaches to personnel assessment, the members were equally critical of their own work. Concerning the validity of the assessment procedure, the authors concluded that, "the final validity is a question mark" (OSS, 1948, p. 392). MacKinnon (1977) recalled that the assessment staff at OSS was forced to conclude that they had failed in predicting performance overseas. It is interesting therefore to analyze what contribution the WWII assessment procedures made to personnel psychology. In his review of the book _Assessment of Men_, Donald E. Super reported in the _Journal of Applied Psychology_ that the psychologists approached their assignment with freshness, originality, and "unusual creativity" (Super, 1949). On the other hand, Super noted that many of the mistakes made by the assessors could have been avoided had they been more familiar with the principles and practices of personnel psychology. Nevertheless, Super's own evaluation of the program was more positive than that of the OSS assessment group. He noted that the average validity of the two principal assessment stations (i.e., the 3-day assessments at Station S, and a 1-day assessment in Station W) for predicting performance ratings of men sent overseas was .45—hardly negligible to a personnel psychologist. Super encouraged businesses to exploit the possibilities of the OSS approach. Less sanguine was the review published in _Personnel Psychology_ by Harry J. Older (1948). Older noted that, although persons not trained in critical analysis will be readily "sold" by the procedures described in _Assessment of Men_, he failed to find conclusive evidence for the utility of the new approach. According to Older, the lack of data on the validity of the separate procedures and the failure to provide information on the correlation of the dimensions with the overall rating made the incremental value of the holistic approach indeterminable. Eysenck (1953) expressed similar concern over the failure to validate procedures separately. According to Eysenck, "The fact that it was not done indicates

---

3 Although former candidates had speculated about other unusual procedures such as planting undercover assessors disguised as candidates (MacDonald, 1947), or blowing hot air into the room during the debates (Hunt, 1974), I was unable to find corroborative evidence to support these claims.
impatience with scientific proof and the sovereign disregard of patient verification of theories, which is so characteristic of the intuitive holistic school" (1953, p. 146). Particularly troubling to both Eysenck and Older was the fact that the 1-day assessment had considerably higher validity (.53) than the 3-day assessment (.37), suggesting that the additional opportunity for clinical judgment only detracted from the validity of the more standardized procedures.

*The Whole Picture*

Despite criticism from some quarters, the holistic approach advocated by the WWII psychologists appeared to be having an impact on executive selection following the war. Procedures adopted from the WOSBs were being implemented in industry in England (e.g., Frazer, 1947; Trist & H. Murray, 1990), and Taft (1948) described the use of similar approaches for the selection of trainee executives in Australia. Earlier, Taft (1946) noted that, in the United States, there was movement away from “mass psychometrics” to a more clinical approach that takes into account the individual as an “organic whole,” whose behavior is the result of personality and environment. A decade later, Stagner (1957) noted that simple, straightforward tests of intelligence and other abilities do not have much value for executive selection because a person is typically not considered for such a position until these aptitudes have been demonstrated. Stagner noted that projective tests and clinical interviewing played increasingly important roles in selection at the higher levels.

Although projective techniques were in existence prior to the war, their entry into the workplace occurred as a result of the perceived success of the WWII approaches and the emergent clinical approach to dealing with organizational problems (Brower & Weider, 1950). According to Abt and Bellak (1950), “The fact that projective methods have entered the fields of social and industrial psychology is evidence of the profound changes that have taken place in recent years in the entire field of psychology, movements that reflect a change in focus from ‘segmental man’ to ‘total man’ ” (p. 435). The use of a more holistic approach to assessment was in keeping with the human relations movement that was gaining considerable momentum in psychology applied to the workplace. For example, it was not unusual in the 1940s and 1950s for organizations to have trained therapists conducting counseling sessions with employees during working hours (see Highhouse, 1999).

One projective device that was specifically designed to explore reaction tendencies to work situations was the Tomkins-Horn Picture Arrangement Test (PAT; Tomkins & Miner, 1957). The PAT consisted of
a number of situations with pictures arranged in random sequence (see Figure 1). The candidate’s task was to arrange the sequence and write a short sentence after each. Three sample responses for the situation in Figure 1 were (Brower & Weider, 1950; p. 454):

△ “The foreman introduces the new worker to his work bench.”
□ “He gives the worker a word of encouragement on starting his job.”
○ “The new worker does his job with gusto.”

From such responses, inferences would be drawn about the candidate’s tendency to assume leadership, degree of dependence on others, perfectionism, persistence, and reaction to criticism. Brower and Weider reported using the PAT on over 2,500 applicants, but cautioned that top-level executives tended to view such figures with “great distance and even disdain” (1950, p. 451). Miner and Culver (1955) used the PAT to identify differences between samples of executives and professors. The authors found that executives more frequently described the hero as experiencing anxiety over illness and more frequently indicated that the worker required assistance in solving a problem. Miner and Culver concluded, “the typical executive is a person who suffers from fears of failure and illness and who has a rather deep conviction of his own helplessness in attempting to solve many of the complex problems that face him” (1955, p. 352). Despite assertions of the predictive validity of the PAT (e.g., Miner, 1961), the measure was criticized for its complex scoring system and its tendency to provide the administrator with a “spurious sense of objectivity” (Semeonoff, 1976; p. 288). The Tomkins-Horn PAT is no longer commercially available.4

Following his work with the PAT, Miner (1965) went on to develop a sentence completion scale for assessing the motivation to manage. The Miner Sentence Completion Scale (MSCS) contains 40 open-ended items (e.g., My family doctor…) aimed at assessing attitudes toward authority, competitiveness, masculinity, desire to stand out, imposing wishes, and liking for routine administrative functions. Like the PAT, scoring of the MSCS involved paying particular attention to rare responses. Miner (1977) summarized a number of studies showing relations between the MSCS and management success indicators. Despite the popularity of the MSCS in the 1970s, Ryan and Sackett (1992) reported that only a small proportion of assessors (less than 10%) continue to use it.

4 A related test using stick figures to represent Murray's need variables (Forrest & Lee, 1962) is evidently still in use, as one outplaced executive recently recounted to me that he was presented (only) one of the figures by an assessor and was asked to write an essay.
Figure 1: Example Situation from the Tomkins-Horn Picture Arrangement Test.

Other projective methods commonly used in industry were word-association techniques (e.g., The Cornell Word Form), the Draw-a-Person, the TAT, and the Rorschach. The latter two techniques were typically reserved for use in assessing candidates at the executive levels (Brower & Weider, 1950). Although various anecdotal records suggested that the Rorschach was being used by consulting firms (e.g., Wickert & McFarland, 1963), very little research exists on its application to industrial settings (cf. Piotrowski & Rock, 1963). The TAT, on the other hand, was getting attention from both consultants and researchers (McClelland, Atkinson, Clark, & Lowell, 1953). Particular attention was given by McClelland and his associates to the assessment of three of Murray’s (1938) needs—achievement, affiliation, and power. These psychologists developed an achievement motive scoring system for the TAT, and founded the consulting firm McBer and Company (now a division of Hay Associates) to assess and train executives on the need for achievement motives. Although personnel psychologists rarely use the instrument today, Ryan and Sackett (1992) reported that nearly one of three psychologists trained in other areas currently use the TAT for executive assessment.

Postwar Programs

The most well documented postwar program combining clinical with psychometric approaches to individual assessment was conducted in the 1950s by Psychological Research Services of Western (now Case Western Reserve University (e.g., Campbell, Otis, Liske, & Prien, 1962; Huse, 1962). The typical procedure used by this group was to visit a client organization to obtain information about the job, the structure and “personality” of the company, and the social environment. This information formed the basis for selecting the battery of tests that would be used in assessment. Tests might include an intelligence test, paper-and-pencil personality instruments (e.g., the Guilford-Zimmerman), and projective devices (e.g., TAT). Two psychologists interviewed the candidate, and a clinical psychologist examined responses to the projectives. A report writer then assimilated this information in the form of a written report for the client organization. Followup research indicated that use of the overall report resulted in lower validity (.11) than use of either the objective tests (.28) or projective devices (.21) alone.

Katzell (1957) acknowledged that the clinical approach to prediction was gaining increasing prominence in industry for the selection of personnel psychologists. McClelland only became interested in personality and the TAT after he had accepted an assignment at Bryn Mawr College to replace Donald MacKinnon, who had departed to work on the OSS assessment project.
executives, but that very little of it was appearing in print. Although he encouraged more validation of research using the clinical approach, KatzeU commented: "The time honored actuarial or statistical approach continues, happily, to dominate the psychological literature on industrial personnel selection" (p. 256). Actuarial approaches to executive assessment were continuing to garner the most attention from personnel psychologists largely because these approaches were highly amenable to research, analysis, and dissemination. Two of the model programs conducted following the war took place at Standard Oil of New Jersey (now Exxon) and Sears, Roebuck and Company. The program at Standard Oil was developed in the early 1950s in response to senior management's concern that the pool of young managerial talent had been depleted by the war and disruptions caused by earlier reorganizations. Prior to this time, the company had relied on a traditional succession system, based on armchair identification and grooming of young stars. The aim of the program was to make valid predictions on the basis of cognitive and noncognitive tests, not to generate "descriptive profiles full of psychological meaning but devoid of criterion relationships" (Sparks, 1990, p. 103). The test battery was administered to existing managers and included measures of verbal ability, inductive reasoning, situational judgment, management attitudes, and personality (e.g., the Guilford-Zimmerman Temperament Survey). In addition, managers completed a biographical questionnaire and a picture test similar to the TAT. Responses to the picture test were scored using an a priori scoring system. A "success index," used as the criterion, was composed of position level, managerial effectiveness (a ranking by higher level executives), and salary. The effects of age and years of service were statistically controlled. The results revealed that, with the exception of the picture test, all of the measures were valid predictors of managerial success. The multiple correlation was an impressive .70. Nevertheless, interest in the entire project began to wane by 1980, and the program was abandoned. According to Sparks (1990), corporate management reinstated the succession system that had been used up until 1953.

In 1942, Sears, Roebuck and Company sought the assistance of L.L. Thurstone in establishing a psychological testing program for executive personnel. The test battery consisted of the American Council on Education Psychological Test, the Guilford-Martin Inventories, The Allport-Vernon Survey of Values, and the Kuder Preference Record. Multiple correlations reported by Bentz (1967) were often as high as .75. Bentz argued that there is a cluster of executive psychological characteristics that transcend boundaries of specialization. Indeed, the notion of a g factor of executive qualities, comprised of forcefulness, dominance, assertiveness, and confidence, has received considerable empirical support (e.g.,

The programs at Sears and Standard Oil were not highly representative of the state of practice in individual assessment during the 1950s and 1960s. Even the more clinically oriented approach of the Western Reserve Group was probably on the high end of rigor, considering that it was published in the journal *Personnel Psychology*. It is likely that procedures loosely described in more practitioner-oriented journals (e.g., Phelan & Smith, 1958) were more representative of practices for assessing higher level personnel. For example, Campbell et al. (1970) interviewed selection specialists from 33 firms widely regarded to be at the forefront of research on managerial effectiveness in the 1960s, and found that all of them used only clinical combinations of their selection information. Furthermore, Dunnette (1967) noted that nearly all consulting firms were relying on psychological appraisals as an aid in making personnel decisions at high levels. For example, Rohrer, Hibler, and Replogle, a large and well established provider of psychological assessment services to industry, relied almost exclusively upon judgments by clinical psychologists using diagnostic interviews and projective devices (RHR International, 1991).

**Brain Watching Under Siege**

Another reason for the inattention to actuarial programs was likely the increasingly negative attitude toward paper-and-pencil psychological testing in the United States. W. R. G. Bender of DuPont Company argued that psychological testing grew out of an outmoded, elementalistic view that an understanding of man could be achieved by adding up measurements of individual parts (Bender, 1964). In addition, Korman (1968) published an influential review of the research on executive assessment, arguing that the prevailing negative view of judgmental (clinical) prediction held by personnel psychologists was simply wrong. Korman concluded that judgmental prediction methods, particularly those used in executive assessment, were generally better predictors than psychometric procedures. According to the author, “all in all then there is little reason for anyone to be contemptuous of ‘judgmental prediction’ in management selection, an attitude that is undoubtedly dominant in industrial psychology” (Korman, 1968, p. 312). In reviewing the state

---

6Korman's comparison of clinical and actuarial judgment was based on comparing composite clinical judgments (based on multiple predictors) with the validity of individual paper-and-pencil predictors—a seemingly unfair comparison. He also classified the work at AT&T as clinical, even though they abandoned many of the clinical portions of their assessment center.
of personnel selection, Guion noted a change in attitude and predisposition toward more concern for the "wholeness and integrity of applicants" (Guion, 1967, p. 105). According to Guion, this change was likely spurred by critics of testing (e.g., Gross, 1962), the civil rights movement, and increased concern for human values.

In 1965, the United States Congress held hearings on the use of tests in employee hiring by the government (see Amrine, 1965). Led by Senator Sam Erwin and Representative Cornelius Gallagher, Congress was interested in the invasion of privacy resulting from the use of tests such as the Minnesota Multiphasic Personality Inventory (MMPI) for employee screening. During the 1950s, it was not unusual for a psychologist to use the MMPI in the assessment of executives. In subsequent years, it had become common practice for personnel administrators to use this test for hiring at other levels (Hedlund, 1965; Matarazzo, 1990). Nevertheless, George Bennett, president of Psychological Corporation, testified before Congress that the MMPI was not a test to determine who should be employed in a job. According to Bennett, "the task of selection must be performed by skilled, trained, and conscientious men and women using the best tools available. The MMPI is not such a tool" (Testimony before the Senate Subcommittee, 1965, p. 941). Despite the fact that the MMPI was the focus of the congressional scrutiny, the taint spread to all paper-and-pencil personality tests used for employment. In fact, the use of the term "personality test" became synonymous with the use of the MMPI (Haney, 1981). In his book The Brain Watchers, Gross (1962) made little distinction between tests used to identify psychopathology and tests used to identify normal personality traits. Furthermore, Gross was highly critical of the use of testing for identifying executives, and he glorified executive recruiters who rejected psychological testing. Such criticisms in the popular press, together with unflattering critiques of personality testing in the personnel psychology literature (e.g., Guion & Gottier, 1965; Korman, 1968), led to a virtual moratorium on paper-and-pencil personality testing for selection that lasted for nearly 2 decades.7

The Assessment Center Movement

Although I have argued that the philosophy and activity of individual assessment was directly influenced by the assessment programs of WWII, those early programs were also the impetus for another movement in managerial assessment—the assessment center. It would be unproductive to argue about which movement (i.e., individual assessment or the assessment center) was more influenced by the war activities, as both

---

7 Contributing to this was the explicit reference to "ability tests" in the Tower Amendment to Title VII of the Civil Rights Act.
were influenced in different ways. (Indeed, the WWII assessment activities also influenced the development of other innovations, such as team building and sensitivity groups.) Because the popularity of the assessment center movement dwarfed that of individual assessment beginning in the late 1960s, it is worthwhile to briefly address some of the features that distinguished the two approaches.

The assessment center grew out of the perceived need to assess non-intellectual aspects of human performance, along with an interest in the situational tests developed by the WOSB and OSS assessors. Unlike the approach of the WWII assessors, however, the situational tests designed by personnel psychologists were aimed at obtaining greater standardization in both testing conditions and scoring (Flanagan, 1954). Early situation tests contained a score sheet of behaviors that the examiner merely checked off during observation. A score was obtained by simply counting the various actions that were previously identified as indicating a behavior of interest. This was clearly seen as a modification of the OSS holistic assessment approach. Moreover, unlike individual assessment's reliance on individual expert judgment, the goal of standardization was to make it possible for the tests to be evaluated by "relatively untrained types of personnel" (Flanagan, 1954, p. 463).

Although the term assessment center was first used by Super (1949) in describing the OSS assessment program, it did not become a part of the personnel psychology lexicon until the publication of a report on the innovative Management Progress Study conducted by AT&T (Bray & Grant, 1966). Ironically, the original AT&T assessment center utilized a more clinical approach than many individual assessment programs. Briefly, the Management Progress Study involved 3½ days of psychological assessments, including projective tests, clinical interviews, work samples (e.g., an in-basket test), group problems, and leaderless group discussions. A typical assessment first involved reading a candidate's autobiographical essay, considering reports on the tests and interviews, and reading Q-sort items the candidate selected as "most" or "least" like him. Staff members rated the candidate on 25 characteristics and then evaluated the candidate's overall potential as a management person in the Bell System (Bray & Grant, 1966). The clinical nature of this early assessment center may not be so surprising, however, considering that Douglas Bray was involved in the OSS assessment program as a graduate student. Nevertheless, because the situational tests were found to account for a large proportion of the variance in the overall judgments of the assessors, the assessment centers were modified for operational use by dropping the use of the clinical interview, Q-sort task, and projectives.

The dominance of the assessment center within personnel psychology was evidenced by the fact that an entire chapter was devoted to it in
the 1976 *Handbook of Industrial and Organizational Psychology* (Finkle, 1976), yet no attention was paid to individual assessment. Moreover, reviews of personnel selection published in the *Annual Review of Psychology* stopped covering executive assessment in the 1970s and 1980s but consistently reported on the popularity of the assessment center. This inattention to individual assessment does not mean that the practice had faded away. By all accounts, individual assessment was alive and well during this period (e.g., Hansen & Conrad, 1991; Jeanneret & Silzer, 1998; Wysocki, 1981). Just as assessment centers thrived during the antitesting movement due to the emphasis on situational assessment, individual assessment survived due to its focus on the clinical interview—a practice that was seemingly immune to the public attacks levied against paper-and-pencil tests. Furthermore, individual assessment was often practiced by psychologists who were trained outside of personnel psychology and who often emerged from clinical psychology programs that housed counseling specialists interested in vocational assessment. An entire field of psychological assessment had emerged as a practice separate from psychological testing (Matarazzo, 1990; Sundberg, Snowden, & Reynolds, 1978). The personality assessment tradition is rooted in the ideas of Henry Murray, whereas the personality testing tradition was influenced by the contrasting work of Gordon Allport (Holtzman, 1964; Winter, John, Stewart, Klohnen, & Duncan, 1998). The evolution of psychological assessment in clinical settings following WWII is remarkably similar to the evolution of individual assessment in industry (see Korchin & Schuldberg, 1981).

Tenopyr and Oeltjen (1982) lamented that the label assessment center was being used indiscriminately to identify practices that varied widely in content and process. However, the use of multiple (nonpsychologist) assessors evaluating multiple candidates, along with the emphasis on situational tests, made assessment centers unique from traditional individual assessment (Finkle, 1976). According to Finkle, most individual assessments conducted by consulting organizations relied on clinical interviews, personality tests, and projectives. Moreover, the evolving trait-focused approach of the assessment center differentiated it from the holistic approach taken by psychologists writing individual assessment reports for organizations. The standardization characteristic of assessment centers made them easier for the hard-nosed testing community to digest (Guion, 1967; Hinrichs, 1978).

---

8Projectives were, however, severely attacked by the testing community. Nevertheless, the Rorschach and TAT were still being used in 91% of clinical settings in 1971 (Rabin, 1981).
The Rediscovery of Individual Assessment

Although individual assessment remained outside of the personnel psychology radar screen for nearly 2 decades, it was “rediscovered” in a series of influential articles by Ryan and Sackett (1987; 1989; 1992). The authors were puzzled by the fact that individual assessment received little research or even description, even though many psychologists were making a living practicing it in organizations (Ryan & Sackett, 1998). More than anything, the work of Ryan and Sackett served to document the practices of modern individual assessors and to define the technique as something different from the assessment center. In one article (Ryan & Sackett, 1992), the authors documented that individual assessment was being widely practiced by psychologists with training in personnel psychology, as well as by psychologists with training outside of the field (e.g., clinicians). The authors also showed that practices often differed by area of training. For example, many more personnel psychologists used situational exercises, while many more nonpersonnel psychologists used the MMPI. Interestingly, a considerable number of both groups still relied upon projective devices, even though these devices had been met with disfavor by the testing community.

The variety of approaches used by modern assessors is evidenced in two books published on individual assessment in the 1990s (Hansen & Conrad, 1991; Jeanneret & Silzer, 1998). The Hansen and Conrad (1991) book was directed more toward the consumer of individual assessment and contained examples of approaches taken by various firms and interviews with assessors who had practiced individual assessment since the 1950s. The view toward validation is best summarized by J.R. Porter, formerly of Psychological Service of Pittsburgh, who commented, “the psychologists who are accurate stay in business” (interviewed in Hansen, 1991). The Jeanneret and Silzer (1998) book was directed more toward personnel psychologists and included discussions of assessor reliability and (predictive) validity, along with an examination of theoretical frameworks.

Individual assessment is no easier to categorize today than it was 50 years ago. Firms such as RHR International (formerly Rohrer, Hibler, and Replogle) still rely heavily on the clinical interview (RHR International, 1991), whereas other firms rely more on structured approaches (e.g., Arnold & Kruse, 1991; Baehr, 1991). Levinson (1998) noted that the holistic view of personality is alive and well in individual assessment, and argues that the “objective” approach to assessment fails to provide a picture of the total person (see also Yeager & Brenner, 1994). Clearly, there is little more unanimity of opinion on individual assessment than existed in the past (Ryan & Sackett, 1998). If there is any agreement,
though, it is on the notion that executive selection should differ from selection at other organizational levels. As Hansen and Conrad (1991) noted, an executive is unlikely to fail because of technical mistakes,

Rather, a person is more likely to get fired because someone in a position of influence thinks the employee is a 'jerk,' or lacks one of a number of vital personal qualities (e.g., 'he has no guts'; 'she's not a likeable person'; 'he's a pompous windbag'; 'he doesn't know how to put his best foot forward').

(p. xii)

Thus, the goal of the typical individual assessor is not to represent the whole personality of a candidate, but the candidate's "personality as a whole" (OSS, 1948). Below, I consider the practices and assumptions of individual assessment, particularly the holistic view of assessment, in light of the current state of knowledge on judgment and prediction.

Assessing Individual Assessment

According to the OSS report (1948), "Organismic [holistic] assessment is based on the hypothesis that a trained psychologist or psychiatrist, with a fund of additional facts at his disposal, is today capable of improving to a significant degree the accuracy of mechanical predictions derived from test scores alone" (pp. 51-52). It is clear from reading the OSS report that this assertion referred not only to incorporating clinical assessments into a final decision, but also to using clinical intuition to combine an array of test scores and personal observations into a diagnostic judgment of future success. The notion that clinical expertise is necessary for combining predictors and making inferences about executive success is the crux of the distinction between the holistic school and the psychometric approach. Modern individual assessment practice continues to reflect a belief in the holistic approach to assessment despite a lack of evidence for its efficacy.

Predicting executive success. The controversy over the use of clinical versus psychometric judgment began in earnest following the publication of Meehl's (1954) influential book, Clinical Versus Statistical Prediction. Meehl argued that most clinical judgments are best made statistically, not intuitively. In support of this argument, Goldberg (1965) showed that a simple linear composite of five MMPI scales outperformed 13 clinicians in predicting a dichotomous diagnosis. In a review of 45 studies comparing clinical to formula prediction, Sawyer (1966) found none in which clinical prediction excelled. Even unsophisticated models usually outperform experts in making predictions (Dawes & Corrigan, 1974; Ree, Carretta, & Earles, 1998; Wainer, 1976). Meehl (1986) noted about the clinical versus statistical prediction controversy: "There is no
controversy in social science that shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one" (pp. 373–374). This is not to say that clinical inference has no place in the assessment process, an issue I will return to later, only that such assessment is best codified and mechanically integrated with other observations.

Individual assessors rarely have the luxury of incorporating regression models derived from predictor-criterion relations. In the absence of regression weights, however, there are several methods of mechanically combining assessment information using "improper" models. For example, one might derive implicit weights from overall judgments using a bootstrapping method (Bowman, 1963). Other simpler options include using rationally derived weights based on judgments of cue importance, or simply weighting all of the cues equally. All methods have been shown to improve prediction over holistic judgments (Dawes & Corrigan, 1974). Unfortunately, one of the likely reasons for the resistance to mechanical combination is also one of the reasons for the poorer performance of clinical judgments in prediction. This is the tendency for experts to use configural rules in making judgments (see Camerer and Johnson, 1991, for an excellent discussion of other reasons for the poor performance of clinical judgment). With configural rules, the impact of one variable depends upon the values of other variables. Dunnette (1967) expressed confidence that subgrouping analysis methods, such as configural scoring or actuarial pattern analysis, would lead to greatly increased levels of validity in prediction of managerial success. In fact, there has long been a feeling among many in the personnel psychology community that more success could be achieved in managerial prediction by exploring configurations of traits that lead to success (e.g., executive styles; see Wickert & McFarland, 1967). More recently, Hogan, Hogan, and Roberts (1996) cautioned against analyzing personality dimensions individually because the manner in which any individual trait operates depends on the full constellation of personality characteristics (see Allport, 1962, for a similar argument). To date, however, little evidence exists to show that nonlinear methods add to prediction (Abrahams & Alf, 1972; Mount, Barrick, & Straus, 1999; Zedeck, 1971). Certainly, we should continue to search for meaningful configurations of personality traits considering, for example, that predictive accuracy may be higher for nonlinear models even when linear models account for more variance in the criterion (see Guion, 1998). Camerer and Johnson (1991) noted, however, that configural rules are highly alluring to assessors because it is easy to weave a causal narrative in trying to explain past cases that are unlikely to replicate (e.g., "We have not had success in hiring executives from Big 10 schools who have previously worked in..."
consulting firms"). Camerer and Johnson further suggested that, rather than abandoning such rules when they prove wrong, people have a tendency to refine them further.

The use of a mechanical method for combining assessment cues does not preclude the use of clinical judgment in the assessment process (Highhouse, 1997). For example, Meehl (1954) acknowledged that humans can excel over formulas in selected predictive tasks. As an example, he proposed the case of Professor X, for whom a statistical equation would predict a .90 probability of going to the movie on a particular night. If, however, one has knowledge that Professor X recently broke his leg, one could predict that he would not go to the movie on a particular night—the equation would have failed to make such a prediction. This has come to be known as the "broken-leg" cue in clinical judgment. These are cues that are rare but highly diagnostic allowing, for example, financial analysts to outpredict novices because the experts can interpret the impact of news events (Johnson, 1988). It is perhaps the ability to recognize broken-leg cues that has produced a series of results showing that combinations of intuitive and mechanical judgments can outperform mechanical judgments alone (e.g., Blattberg & Hoch, 1990; Whitecotton, Sanders, & Norris, 1998). Most relevant to the assessment context was research by Ganzach, Kluger, and Klayman (2000) showing that expert "fine tuning" of mechanical combinations of employment interview scores resulted in more accurate predictions than when the mechanically combined scores were used alone. Of course, systematically adding clinical judgment to a formula or fine tuning formula-based predictions is far different from intuitively deriving an overall judgment. Existing evidence, however, suggests that standard practice in individual assessment is to integrate cues using an overall clinical judgment about candidate effectiveness (e.g., Campbell et al., 1970; Ryan & Sackett, 1998).

The problem with relying on holistic judgments is determining which cues are broken-leg and which are merely error. (Is identifying someone as a "windbag" a broken-leg cue?) Moreover, when assessors intuit a theory of the person being assessed, they often rely too much on unreliable cues or combinations of cues. This is likely due to the use of heuristics such as representativeness and availability (Highhouse, 1997) and a tendency to cling to the belief that humans can become intuitive experts in predicting future behavior (Kleinmuntz, 1990). Yet, the relation between clinical experience and accuracy in judging personality is near zero (Garb, 1989). In a study of expert judgments, Camerer (1981) found that 60% of the variance in residuals was systematic use of configurual rules and broken-leg cues. The correlation between residuals and outcomes, however, averaged only about .05. Indeed, the notion of
expertise in prediction has been shown to be an illusion for occupations as diverse as psychologists, doctors, accountants, and parole officers.

*Validating inferences.* All of this discussion may be moot if the psychologist does not offer a hiring recommendation. Many individual assessors merely offer psychological profiles, which the client organizations may either use or disregard (Ryan & Sackett, 1992). In fact, Guion (1998) argued that managers, as consumers of assessment information, should not surrender the hiring decision to individual assessors. According to Guion,

> In my own (individual assessment) work, I explicitly refused to make recommendations, preferring to end appraisal reports with a series of questions suggested by my conclusions about the candidate. These questions were intended to force the person ultimately responsible for a hiring decision to get independent information... and to form his or her own conclusions without abdicating responsibility to an outsider. (p. 633)

This shifts the validity focus away from prediction, toward the construct validity of inferences made from observations of candidates. If, for example, the psychologist reports that the executive candidate is a “jerk,” then it is important that the observed behavior is indicative of the jerk construct. Some progress along these lines has been reported by Hogan et al. (1990), who have identified characteristics of “flawed” managers. Their inventory of personal motives is designed to assess negative personal qualities, such as narcissism, that tend to derail executives. Hogan et al. report positive relations between scores on the measure and independent measures of personality disorders. However, such validation of inferences has been the exception rather than the rule in individual assessment. Little mention is ever made of content validation in the practice of individual assessment, suggesting a belief that the professional guidelines regarding content validity evidence are not relevant for higher level assessment (Ryan & Sackett, 1992). As Guion (1998) has noted with regard to personality testing, content validation involves simply: (a) identifying the domain of interest (job), (b) identifying relevant behaviors in that domain, and (c) identifying traits that reflect consistencies in those behaviors. This is clearly an approach that can be utilized by assessors in developing an individual assessment program. Unfortunately, a lack of focus on the job at hand is one of the reasons that more variance is accounted for by assessor differences than by applicant differences when several assessors make judgments about the same set of applicants (Ryan & Sackett, 1989). Too often, assessors base clinical judgments on a comparison of the candidate with a general profile of an “effective executive,” although having no familiarity with the job under consideration (Gatewood & Field, 2001).
Because a considerable minority of individual assessors continue to use projective devices (Ryan & Sackett, 1992), it seems especially important to consider the status of these devices in the scientific literature. The extensive research on graphology has shown the validity of handwriting analysis to be negligible (Dean, 1992). The status of the Rorschach continues to be a lively topic of debate, ranging from those who conclude that the procedure provides almost no incremental validity over other assessment information (Lilienfeld, Wood, & Garb, 2000) to those who argue that the research base supporting the Rorschach has been overlooked (Viglione & Hilsenroth, 2001). Not debatable, however, is the absence of evidence showing that the Rorschach is useful for making inferences about occupational success. Hunsley and Bailey (2001) concluded, “there is so little replicated evidence stemming from high quality studies that we believe there is no scientific basis to support the continued widespread use of the Rorschach in clinical, forensic, and occupational settings” (p. 483). A recent meta-analysis of studies using Miner’s Sentence Completion Scale showed that MSCS scores correlate (corrected r = .35) with performance ratings, and that subscale scores showed convergent and discriminant validity in predicting scores on a Big Five measure (Carson & Gilliard, 1993). Because this meta-analysis was performed on a small number of studies, a majority of which were conducted by Miner and his students in the 1970s, it is unclear whether the MSCS items (e.g., Playing golf..., Punishing children...) continue to be relevant or whether the scoring system can be used by people not trained in Miner’s theory (see Brief, Aldag, & Chacko, 1977).

Despite the failure of the TAT-Based Picture Test in the studies conducted at Standard Oil of New Jersey (Sparks, 1990), some evidence has been found for the validity of the TAT as a measure of motives (Spangler, 1992). Spangler conducted a meta-analysis of 105 studies on the relationship between the TAT and self-reported indexes of achievement motivation. The average relation between these measures and occupational success measures was .22 for the TAT and .13 for the self-report measures. Although the effect size was small, it provides some support for the notion that the measurement of motives may provide some validity over the measurement of traits (the mean correlation between the TAT-based achievement motive ratings and self-report achievement motives was .09). Indeed, some research has shown that the interaction of traits and motives (e.g., Introversion and affiliation motive) predict life outcomes better than either one alone (Winter et al., 1998). It should be noted that Lilienfeld et al. (2000) found adequate validity evidence for only a few of the many scoring schemes for the TAT. Even more troubling is the finding that only 3% of psychologists using the TAT rely on any kind of standardized scoring key (Pinkerman, Haynes, & Keiser, 1993). The
vast majority of psychologists interpret the TAT on the basis of clinical judgment and intuition.

**Functional Autonomy?**

The primary objectives of this article were to establish the place of individual assessment in the history of personnel selection, and to provide a critique of the practice in view of what we have learned about psychological decision making over the last 50 years. The article has shown how the holistic view of personality, most notably associated with the work of Henry Murray, was dropped by the followers of the assessment center movement but continued as a guiding principle in individual assessment. This, combined with the particularly thorny small-sample problem and criterion problem in individual assessment, has kept the practice outside of the mainstream of personnel psychology (see Austin & Villanova, 1992, for an historical discussion of the criterion problem in applied psychology). The commonly held belief that executives cannot be summed up by their scores on standard paper-and-pencil tests, together with the public displeasure with testing in the 1960s and 1970s, has allowed individual assessment to survive and even thrive as an area of psychological practice in organizations. Moreover, recent surveys of individual assessment practices have institutionalized the practice as an independent area of professional practice within personnel psychology (Jeanneret & Silzer, 1998; Ryan & Sackett, 1987, 1989, 1992).

The sample size problem in individual assessment can be partially addressed if consulting firms are willing to share data on a large number of assessments conducted across organizations (Ryan & Sackett, 1998). However, such aggregation invites a host of other issues such as interassessor (in)consistency and criterion contamination. Ryan and Sackett note that such obstacles can only be truly overcome by the unlikely sponsorship of a “research only” assessment program. Moreover, it is unlikely that the criterion problem in individual assessment will ever reach the degree of resolution that has been reached for criteria at lower levels (Hulin, 1963). Executives are blamed for poor performance more often than they deserve and are given credit for good performance more often than they deserve. One need only look at the revolving door of coaches in professional sports to see evidence for both of these propositions. Indeed, even with the benefit of hindsight, historians frequently disagree on the performance of people at the highest executive office. The rediscovery of individual assessment by personnel psychology should, however, lead to increased attention toward investigating the validity of inferences made on the basis of behavioral observations, responses to projective devices, and constellations of personality traits. Increased
attention also needs to be directed toward identifying the criterion space of executive performance. Such work is critical for the art of individual assessment to be reconciled with the science of personnel decision making.

As this review has shown, the holistic approach to judgment and prediction has simply not held up to scientific scrutiny. Very little research has been conducted on the efficacy of individual assessment practices, and the research base that does exist is vaguely described and outdated. As such, it is puzzling why individual assessment has not been subjected to scientific investigation in the way that other selection practices have. In discussing how psychotherapy had flourished in the 1950s despite the unimpressive evidence for its efficacy, Astin (1961) concluded that the practice had achieved *functional autonomy*. The same might be said for individual psychological assessment. Workshops on the practice continue to overflow at conferences for the Society for Industrial and Organizational Psychologists, and the practice continues to expand into other areas, such as executive development and coaching (Jeanneret & Silzer, 1998). Practitioners with no training in personnel selection and EEO issues continue to flood into this area without anyone challenging them to provide professionally acceptable evidence for the veracity of their claims. Similar to psychotherapy, it seems that the principle of functional autonomy has enabled individual psychological assessment to survive and flourish. Now seems the time to take a step back and call for some accountability in this area of professional practice.

REFERENCES


Ansbacher HL. (1941b). Murray's and Simoenet's (German military) methods of personality study. *Journal of Abnormal Psychology, 36*, 589–592.


Sutherland JD, Fitzpatrick GA. (1945). Some approaches to group problems in the British Army. *Sociometry, 8,* 205–217.


