Projective Measures of Personality and Psychopathology
How Well Do They Work?

Although projective techniques remain among the most popular of all psychological measures, many critics have branded them pseudoscientific.

SCOTT O. LILIENFELD

Whatever their theoretical differences, all psychologists agree on one maxim: individuals differ greatly from one another in their mental attributes. In the domains of personality and psychopathology, two types of tests have been used to assess these individual differences, objective and projective. Objective measures contain relatively clear-cut stimuli (e.g., such items as "I like to go to parties") and require subjects to respond in one of a few fixed ways (e.g., by answering true or false). Many objective tests, including the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway and McKinley 1940) and its revision, the MMPI-2, have been shown to be useful in assessing both personality and mental illness. For example,
the scales of the MMPI and MMPI-2 (e.g., schizophrenia, depression) are associated with other indicators of psychopathology, such as diagnoses obtained by clinical interviews (Zalewski and Gottesman 1991).

In contrast to objective tests, projective tests typically present subjects with an ambiguous stimulus, such as an inkblot, and ask them to make sense of this stimulus. Unlike objective measures, projective measures grant subjects considerable leeway in their responses, and the range of possible answers is often limited only by the respondent's ingenuity, loquacity, or both. Nevertheless, as Meehl (1945) observed, these two types of tests probably differ more in degree than kind. Objective test items (e.g., "I sometimes feel sad") often entail a certain degree of ambiguity, because they depend partly on subjects' interpretations of subjective terms (e.g., "sometimes").

The rationale underlying projective tests is the projective hypothesis: when interpreting an ambiguous stimulus, subjects ostensibly project aspects of their personality onto the stimulus. The concept of projection derives from Freud (1911), who conceptualized it as a defense mechanism by which individuals unconsciously ascribe their negative attributes to others. Freud's concept of projection ("classical projection"), however, is believed by psychoanalysts to differ from the "assimilative projection" that presumably occurs in projective tests, which is the more general tendency of individuals' personality characteristics to influence their interpretation of ambiguous stimuli (Sundberg 1977). Projective tests are often viewed as the "stealth weapons" of the psychologist's armamentarium. By circumventing respondents' conscious defenses, they purportedly provide information concerning underlying conflicts.

Many projective measures, particularly inkblot tests, capitalize on the phenomenon of pareidolia (Schick and Vaughn 1995), the perception of meaning in visually meaningless stimuli. Although pareidolia can lead to bizarre misperceptions, such as the "Face on Mars" (Gardner 1988), it is probably a byproduct of a generally adaptive propensity toward pattern recognition (Gilovich 1991).

Most projective tests are highly controversial (Gittelman Klein 1986; Lowenstein 1987), largely because the evidence for their reliability and validity is weak. Reliability refers to consistency in measurement; a reliable test yields comparable scores across different occasions (test-retest reliability) and different examiners or scorers (inter-rater reliability). Validity refers to truth in advertising; a valid test measures what it purports to measure. The relation between reliability and validity is asymmetrical. A reliable test is not necessarily valid—imagine a test that purports to assess individuals' intelligence by measuring their heights—but a valid test is necessarily reliable.

Projective Tests: A Brief History

The term "projective test" was coined by Frank (1948), who likened projective tests to psychological "X-rays" that yield fleeting glimpses into otherwise unobservable mental processes. The first projective test was probably devised in 1879 by Sir Francis Galton. Galton's test was a measure of word association: subjects were presented with a set of words and asked to provide their first response to each word. Carl Jung, the Swiss psychiatrist and protégé of Freud, later used the word association test in conjunction with blood pressure indices to detect what he termed "complexes," i.e., constella-

Most projective tests are highly controversial largely because the evidence for their reliability and validity is weak.

Not until the fourth and fifth decades of the twentieth century, however, did projective tests, particularly the Rorschach Inkblot Test, come into widespread use. The aftermath of World War II witnessed a heightened demand for the diagnosis of veterans returning from combat duty with emotional problems. Projective tests thus became a staple in many assessment batteries. Over the past several decades, however, projective tests have become targets of increasing criticism. As noted by Aiken (1996), even advocates of these tests acknowledge that they are often (1) insufficiently reliable or valid to justify their clinical use, (2) easily faked in either a good or bad direction, (3) susceptible to subtle situational factors, including the personality of the examiner and mood of the examinee, and (4) subjective in scoring and interpretation.

A sampling of quotations from research psychologists over the past four decades reveals a persistent dissatisfaction with the status of many projective tests, and the Rorschach in particular:

- "... the rate of scientific progress in clinical psychology might well be measured by the speed and thoroughness with which it gets over the Rorschach" (Jensen 1965, p. 238).

Scott O. Lilienfeld is an assistant professor in the Department of Psychology at Emory University, Atlanta, GA 30322. His interests include the critical evaluation of pseudoscientific and questionable practices in clinical psychology.
• “Projective techniques do not provide an X-ray of the personality nor a ‘royal road to the unconscious,’ as Freud described dreams to be. The Rorschach and other procedures are not magical . . .” (Sundberg 1977, p. 223).

• “Projective techniques present a curious discrepancy between research and practice. When evaluated as psychometric instruments, the large majority make a poor showing. Yet their popularity in clinical use continues unabated” (Anastasi 1982, p. 564).

• “The Rorschach is indeed projective—for the interpreter. Compelling as the interpretations are, and as fun as the test is to give, it doesn’t provide the insight about the subject that its users allege it provides” (Dawes 1994, p. 149).

Nevertheless, projective tests continue to enjoy widespread popularity. Durand, Blanchard, and Mindell (1988) found that 49 percent of directors of clinical psychology graduate programs and 65 percent of directors of clinical psychology internships believed that training in projective techniques was important. Watkins, Campbell, Neiberding, and Hallmark (1995) found that five projective tests, including the Rorschach and Thematic Apperception Test, were among the ten measures most often used by clinical psychologists. For example, 82 percent of clinical psychologists reported that they administered the Rorschach at least “occasionally” in their test batteries and 43 percent reported that they “frequently” or “always” administered it.

A Consumer’s Guide to Projective Tests

The following is a guide to six major projective tests, consisting of a description of each test and a brief evaluation of its psychometric properties (e.g., reliability and validity). Because of space constraints, some of the more unusual or obscure projective techniques (see sidebar) cannot be reviewed in detail here. In addition, the projective interpretation of handwriting (i.e., graphology) has been reviewed elsewhere (Beyerstein and Beyerstein 1992) and will not be discussed here. Suffice it to say that graphology has been found to be of little or no value in assessing personality (Furnham 1988).

(1) Rorschach Inkblot Test. Perhaps the best known of all projective tests was developed by Hermann Rorschach in 1921. The Rorschach consists of ten cards, almost all of which are bilaterally symmetrical. Five of these cards are in black-and-white, and five contain color (see figure 1). The subject is asked to examine each inkblot and to say what it looks like; respondents are typically permitted to provide as many responses to each card as they wish.

Rorschach interpretation involves three sets of variables: location—where on the blot subjects saw what they saw; content—what subjects report seeing; and determinants—what made subjects report what they saw. Determinants, which are emphasized in modern Rorschach interpretation systems, include form, shading, and color. Although numerous Rorschach scoring schemes have been developed (e.g., Beck, 1944), the Comprehensive System of Exner (1991) has gained ascendance during the past decade. Exner’s system, which yields a large number of indices designed to assess numerous psychological characteristics, was developed by retaining the features of extant Rorschach scoring systems that appeared to have adequate research support.

Despite its popularity, the scientific status of the Rorschach has repeatedly been called into question. Wood, Nezworski, and Stejskal (1969) concluded that neither the reliability nor validity of most Exner indices have been satisfactorily established. For instance, although the number of reflection responses (e.g., “A cat looking into a mirror”) is hypothesized to be related to narcissism, the results of several studies reveal no association between reflection responses and narcissism. In addition, Exner’s Depression Index has not been found to be associated with diagnoses of depression. Nor does Exner’s Suicide Constellation appear useful in predicting suicide (Eysenck and Eysenck 1992).

Certain Rorschach indices, however, may possess at least some validity. A quantitative review by Parker, Hanson, and Hunsley (1988) revealed that a number of Rorschach indices possessed validities approximately equal to those of the scales of the MMPI-2. For example, the F+%, which denotes the number of “good form” responses (i.e., responses that correspond to the shape of the inkblot), is inversely associated with schizophrenia and other conditions characterized by disordered thinking. Nevertheless, as Dawes (1994) notes, the F+% and similar indices are actually nonprojective in nature, because they assume that the inkblots do in fact resemble certain stimuli. Consequently, these findings do not provide support for the use of the Rorschach as a projective test, although they suggest that it may be useful for assessing thought disorder. Even here, however, Rorschach appears to be inferior to the MMPI Schizophrenia scale in the prediction of schizophrenia diagnoses (Archer and Gordon 1988).

(2) Thematic Apperception Test (TAT). The TAT, developed in 1935 by Henry Murray and his student Christiana Morgan, consists of thirty-one cards, most of which portray individuals

Figure 1. The Rorschach-like Inkblot: From Anastasi/Uribe, Psychological Testing, 7/E, P. 413. Reprinted by permission of Prentice Hall, Upper Saddle River, New Jersey.
in ambiguous situations (see figure 2). One TAT card, Card 16, represents the epitome of ambiguity; it is entirely blank. The TAT examiner typically selects approximately 10 cards that seem especially relevant to the subject's presenting problems. The subject is then asked to compose a story describing the actions, thoughts, and feelings of the characters on each card. Although the TAT is usually scored impressionistically (i.e., on the basis of unquantified clinical impressions), standardized coding schemes have been developed for scoring certain personality variables. For example, McClelland (1961) created a TAT scoring system to assess achievement needs.

The evidence for the TAT's validity is arguably even more problematic than that of the Rorschach, largely because many TAT studies do not use standardized scoring methods. A review of 105 studies of McClelland's (1961) scoring scheme (Spangler 1992) provided both good and bad news for the TAT. The good news? McClelland's scheme performed about as well as self-report measures of achievement when measured against objective criteria for achievement (e.g., grades). The bad news? Neither type of measure performed especially well; the correlations between objectively assessed achievement and both the TAT and self-report measures were statistically significant (i.e., too large to have arisen by chance), but very low in magnitude.

Many other studies of the TAT's validity have yielded unimpressive results (Gittelman Klein 1986). For example, Sharkey and Ritzler (1985) found that the TAT could not distinguish depressed or psychotic patients from normals on the basis of the emotional tone of stories. In addition, TAT responses have often been found to exhibit poor test-retest reliability (Vane 1981; Winter and Stewart 1977). Because reliability is a prerequisite for validity, this finding does not bode well for the TAT's validity.

(3) Draw-a-Person Test (DAP). The DAP simply asks subjects to draw a person in any way they wish, and is traditionally interpreted in terms of specific bodily signs. For example, large eyes are believed to indicate suspiciousness, and a large head is believed to indicate concerns regarding one's intelligence (Machover 1951).

Nevertheless, research has consistently indicated that these signs are invalid (Kahlil 1984; Swenson 1968). Indeed, the only replicable finding in the DAP literature is that the overall quality of drawing bears a modest relation to the severity of psychopathology (Roback 1968). Gregory (1992) summed up the opinions of many researchers when he concluded that groups was relatively low.

(4) Bender-Gestalt Figure Drawing Test. The Bender-Gestalt asks the subject to copy a series of drawings containing shapes, lines, and dots. Although this test is typically used to assess brain damage, many psychologists have adapted it to assess psychopathology. Certain test behaviors are believed to be indicative of schizophrenia, depression, or anxiety. In a review of the literature on the Bender-Gestalt, however, Naglieri (1992) concluded that there is little support for the use of this measure as a projective device. For example, Trabhan and Stricklin (1979) found that Bender-Gestalt signs purportedly related to aggression (e.g., an increase in the size of the first several drawings, drawings that collide with each other) were not associated with overt hostility. Moreover, there is little evidence that clinical experience enhances the validity of Bender-Gestalt interpretations. Goldberg (1959) found that psychologists were no better than their secretaries (!) at differentiating psychiatric patients from brain-damaged patients on the basis of Bender-Gestalt drawings, and that the diagnostic accuracy of both

(5) Rosenzweig Picture-Frustration Study (PFS). (Rosenzweig, Fleming, and Clark 1947). This measure, which has the distinction of being featured in the film *A Clockwork Orange*, was designed to assess propensities for several types of aggression. It consists of cartoons, each of which contains two people. In each cartoon, the person on the left says something that is potentially anger-provoking, and the subject is asked to state how the person on the right would respond verbally (see figure 3). The evidence for the validity of the Rosenzweig PFS, unlike that for most projective tests, appears promising. For example, scores on this test increase following frustrating situations (Muehlen, Hollinden, and Batsel 1981) and correlate with self-reported (but not teacher-reported) verbal and physical aggression among children (Graybill, Williams, Bodmer, and Peterson 1991). Some of these findings, however, have been inconsistent and difficult to replicate (LaVoie 1986).

(6) Sentence Completion Test (SCT). In this test, of which there are many versions (e.g., Rotter 1946), the subject is pre-
sented with a series of stems (e.g., "If my mother . . .") and is asked to complete each stem by forming a sentence. The evidence for the reliability and validity of most SCT versions is unimpressive (Aiken 1996). One exception is a version of the SCT that has been developed by Loevinger (1976) over several decades.

Loevinger uses her SCT version to assess a construct she calls "ego development," which refers to the maturation of the self. As one progresses from the lowest to the highest levels of ego development, one experiences an increasing capacity to perceive the world in shades of gray. For example, a conformist (low level) response to the stem "When they avoided me . . ." is "I felt as if I wasn't wanted," whereas an autonomous (high level) response is "I wondered why they did—was it me? Or was it them? Or something else?" (Loevinger 1987).

Loevinger's SCT was developed by a sophisticated method of test construction in which the results of SCT protocols were used to progressively refine the scoring key. Although Loevinger's SCT is not without its critics (Snarey, Kohlberg, and Noam 1983), it is one of the few projective tests to demonstrate consistent validity. For example, scores on this test are positively associated with empathy and moral development, and negatively associated with delinquency and psychopathology (Loevinger 1984; Williams and Vincent 1985).

The Bottom Line: How Well Do Projective Tests Work?

The question contained in this article's title is not easily answered in a blanket statement. On the positive side, certain projective tests appear to yield indices with at least some validity. With the possible exception of the Rosenzweig Picture-Frustration Study (PFS) and Loevinger's Sentence Completion Test (SCT), however, the levels of validity are modest at best and nonexistent at worst. It seems safe to conclude that the popularity enjoyed by most projective tests is outstripped by the mediocre, and in most cases feeble, evidence for their validity.

Perhaps the most damaging criticism of most projective tests is their negligible or absent *incremental validity*. The concept of incremental validity, which was introduced by Mehl (1959) and Sechrest (1963), refers to the extent to which a test contributes information to the prediction of a criterion, such as a diagnosis, above and beyond information that is already available (e.g., demographic information, scores on more easily administered psychological tests). There is no convincing evidence that the Rorschach, for example, possesses incremental validity over and above the MMPI or other tests (Wood et al. 1996). An anecdote relayed to me by a former supervisor, which I paraphrase here, nicely illustrates the concept of incremental validity. A graduate student had given the Rorschach to a patient whom he suspected of being homosexual (this incident occurred prior to the mid 1970s, when homosexuality was still considered a mental disorder by the American Psychiatric Association). The student triumphantly reported his findings to his supervisor: "This patient has nearly every homosexual sign on the Rorschach. I was right! As I suspected, he's homosexual." The supervisor looked sternly at the student and replied, "I know. When the perplexed student asked "How did you know?," the supervisor responded, "I asked him."

Another difficulty with the hypotheses derived from many projective tests is their falsifiability. For example, if a Thematic Apperception Test (TAT) protocol reveals a high level of aggression, should the subject be expected to exhibit high levels of aggression in the real world? Or might it instead mean that the subject is projecting aggression that he or she is afraid to express, and should therefore be expected to exhibit *low* levels of aggression in the real world? The latter possibility has sometimes been used to explain away negative findings for the TAT and other projective tests (Vane 1981). This ambiguity makes the findings from many studies of projective tests difficult to interpret.

There is some suggestion that the few promising projective tests, such as the Rosenzweig PFS and Loevinger SCT, may share one feature. Specifically, these tests attempt to assess a single psychological attribute, rather than to provide a comprehensive assessment of many attributes. Why might this difference be important?

With apologies to fans of The Rolling Stones, I propose that the answer may be found in what I like to call "The Rolling Stones Effect." Although some of my friends vigorously disagree with me, I have long maintained that none of the members of the Rolling Stones can sing especially well. Yet when they sing together as a group, they aren't all that bad. The Rolling Stones Effect illustrates what psychologists term the principle of aggregation. By pooling across a number of observations (in this case, singers) the random errors con-
tained in any one observation (in this case, the imperfections in each singer's voice) are canceled out. Psychologists have long recognized that single test responses are unreliable, because each contains a great deal of unsystematic error or "noise" (Epstein 1979), most of which is produced by random situational factors (e.g., the mood of the respondent, the personality of the examiner). By averaging across many test responses, this noise is minimized and the weak "signal" contained in these responses is amplified.

By using a large number of responses to assess a single psychological dimension, the Rosenthal PFS and Loewinger SCT capitalize on the Rolling Stones Effect. In contrast, the Rorschach, which attempts to assess many different traits with only a few responses per trait, ignores this principle. Consequently, the Rorschach, like most projective tests, is a "jack of all trades, master of none." This hypothesis is consistent with the findings of Meyer (1993), who reported that the correlations between several Rorschach indices and MMPI-2 measures of psychopathology were high only when the subject had given a large number of Rorschach responses.

In other words, some Rorschach indices appeared to possess validity, but only when there were enough responses on which to base them.

The open-ended format of projective tests appears to be a double-edged sword. This format permits respondents to structure the test stimuli in any way they wish, thereby allowing personality variables to influence their responses. At the same time, however, the open-ended nature of projective tests permits many extraneous variables—such as the examiner's personality, the specific wording of prompt questions (Aiken 1996), and even the transient motivational state of the examinee—to come into play. Hungry subjects, for example, are especially likely to tell TAT stories involving food (McClelland 1951).

Why Are Projective Tests So Popular?

Given the weak evidence for the reliability and validity of most projective tests, one is left to ponder why they remain so popular. A number of reasons come to mind. First, tradition dies hard. Many clinicians were trained in the use of projective

---

**A Sampling of Other Projective Tests**

In addition to some of the better known projective measures reviewed here, a number of more arcane or bizarre projective tests have been developed. These include the following measures:

- **The Szondi Test** (Szondi, Moser, and Webb 1959) consists of several sets of photographs of actual psychiatric patients suffering from paranoia, mania, and other disorders. Subjects are asked to select the two patients they like best and least from each set, the assumption being that these choices reveal subjects' unconscious needs.

- **The Hand Test** (Wagner 1962) presents subjects with various drawings of moving hands and asks them to guess what each hand "might be doing." The Hand Test uses these guesses to assess aggression, anxiety, and other traits.

- **The Lüscher Color Test**, which became well known following the book by Lüscher and Scott (1969), purports to assess subjects' personality traits by examining their preferences for various colored cards. For example, subjects who prefer blue are hypothesized to possess needs for tranquility; subjects who prefer green are hypothesized to possess needs to impress others.

- **The Blacky Test** (Blum 1950) consists of twelve cartoon drawings of a black puppy engaged in ambiguous activities, and requires subjects to tell a story about each drawing. This test is intended to assess the presence of psychological conflicts derived from Freudian theory. For example, a cartoon of Blacky watching a knife descend upon his sibling's tail is purported to assess castration anxiety.

- **The Mira Myokinetik Psychodiagnostic Test** (Mira 1940) claims to assess psychopathological tendencies by examining the characteristics (e.g., length, variability) of lines drawn by blindfolded subjects.

- **The Lowenfeld Mosaic Test** (Lowenfeld 1949) asks subjects to create mosaic designs out of 465 multicolored wooden shapes, and is purportedly helpful in the diagnosis of schizophrenia, depression, and other conditions.

- **The Draw-A-Car Test** (Loney 1971) asks children to draw a car and asks them twenty-four questions about this drawing (e.g., "How fast will it go?"). It supposedly helps to identify children with encopresis and enuresis (i.e., difficulty in inhibiting defecation and urination, respectively).

- **The Pigem Test** (see Haworth and Rabin 1960) assesses subjects' personality traits by asking them what kind of animal they would like to be if they "had to return to this world and could not be a person."

- **The Rock-A-Bye Baby Test** (Haworth 1961) is a film consisting of puppets enacting the parts of various characters, including a child, a newborn baby, his mother, and an evil witch. It is designed to assess children's jealousy toward younger siblings, guilt regarding their aggressive feelings, and other emotions.

Most of these tests have either been inadequately researched (e.g., the Pigem Test) or found to possess inadequate validity. The Szondi Test, for example, appears essentially useless in the assessment of personality (Cronbach 1960), while the Blacky Test's reliability and validity are both questionable (LaVoie 1984). Moreover, despite its popularity in parts of Europe, the Lüscher Color Test has not withstood careful scrutiny. For example, it has been found to possess low test-retest reliability (Braun and Bonta 1979) and to be unrelated to scores on anxiety measures (McAlonan and Lester 1979) and personality inferences derived from the MMPI (Holmes et al. 1984).
devices and have grown accustomed to administering them for decades. Moreover, many clinicians almost certainly receive communal reinforcement (Carroll 1998) from like-minded colleagues for administering these tests. Second, research on what Meehl (1956) called the P.T. Barnum effect—the tendency of subjects to regard vague and uninformative statements (e.g., “You have a great deal of unused potential”) as accurate self-descriptions—indicates that subjects are especially likely to accept Barnum statements that they believe originated from projective, as opposed to objective, tests (Snyder, Shenkel, and Lowery 1977). Presumably, this is because most subjects view projective tests as mysterious and inscrutable, and leave these tests perplexed about what their responses mean. In the psychotherapeutic context, this greater client acceptance of projective test interpretations may become translated into greater clinician confidence.

Perhaps the principal reason for the popularity of projective tests, however, can be found in the remarkable phenomenon of illusory correlation. This concept was introduced by Loren and Jean Chapman, who demonstrated that individuals often perceive statistical associations between projective text signs and certain personality traits even when these associations are nonexistent. For example, Chapman and Chapman (1967) showed undergraduate subjects a series of fabricated Draw-A-Person (DAP) protocols containing certain physical features (e.g., large eyes, a big head), along with a description of the personality characteristics of the individual who supposedly produced each drawing (e.g., paranoid, concerned about masculinity). Subjects were asked to estimate the extent to which these physical features and personality characteristics co-occurred. Unbeknownst to subjects, there was no correlation between the (DAP) features and personality characteristics, because these two sets of variables had been paired randomly.

Nevertheless, subjects consistently perceived certain DAP features to be associated with certain personality traits. Interestingly, these were the same features that clinicians tend to believe are associated with these personality traits—and that have been found to be invalid in research studies (Kahill 1984). For example, subjects reported that individuals who produced drawings with large eyes tended to be paranoid, and that individuals who produced drawings with large genitals tended to be concerned about their sexuality. In these cases, subjects appeared to rely on their implicit notions regarding the associations between specific physical features and specific personality traits. Chapman and Chapman (1969) later extended these findings to the Rorschach.

The Chapmans’ findings do not demonstrate that projective tests are of no value. But they do suggest that clinicians will often be convinced of the validity of projective tests even in the absence of evidence. Like many “cognitive illusions” (Piantelli-Palmarini 1994), our propensity toward illusory correlation may represent a deeply ingrained by product of the human mind’s propensity to perceive patterns and relationships in random data (Gilovich 1991). If so, clinicians will need to master a skill that does not come naturally to any of us: disregarding the vivid and compelling data of subjective experience in favor of the often dry and impersonal results of objective research.

Acknowledgement
I thank Lori Marino for her helpful comments on an earlier draft of this manuscript.

Notes
1. A third type of reliability, internal consistency, essentially refers to the extent to which the items on a test relate to each other. This type of reliability will not be discussed further here.
2. A recent meta-analysis of Parker et al.’s (1988) data set by Garb, Florio, and Grove (1998), however, suggests that the MMPI is significantly more valid overall than the Rorschach.

References