



BY SCOTT O. LILIENFELD, JAMES M. WOOD AND HOWARD N. GARB

What's Wrong with This

# PICTURE BY PICTURE IN BE E?

PHOTOGRAPHS BY JELLE WAGENAAR

PSYCHOLOGISTS OFTEN USE THE FAMOUS RORSCHACH  
INKBLOT TEST AND RELATED TOOLS TO ASSESS  
PERSONALITY AND MENTAL ILLNESS. BUT RESEARCH  
SAYS THE INSTRUMENTS ARE FREQUENTLY  
INEFFECTIVE FOR THOSE PURPOSES

***What if you were asked to describe images you saw in an inkblot or to invent a story for an ambiguous illustration—say, of a middle-aged man looking away from a woman who was grabbing his arm? To comply, you would draw on your own emotions, experiences, memories and imagination. You would, in short, project yourself into the images. Once you did that, many practicing psychologists would assert, trained evaluators could mine your musings to reach conclusions about your personality traits, unconscious needs and overall mental health.***

But how correct would they be? The answer is important because psychologists frequently apply such “projective” instruments (presenting people with ambiguous images, words or objects) as components of mental assessments, and because the outcomes can profoundly affect the lives of the respondents. The tools often serve, for instance, as aids in diagnosing mental illness, in predicting whether convicts are likely to become violent after being paroled, in evaluating the mental stability of parents engaged in custody battles, and in discerning whether children have been sexually molested.

We recently reviewed a large body of research into how well projective methods work, concentrating on three of the most extensively used and best-studied instruments. Overall our findings are unsettling.

### **Butterflies or Bison?**

**THE FAMOUS RORSCHACH** inkblot test—which asks people to describe what they see in a series of 10 inkblots—is by far the most popular of the projective methods, given to hundreds of thousands, or perhaps millions, of people every year. The research discussed below refers to the modern, rehabilitated version, not to the original construction, introduced in the 1920s by Swiss psychiatrist Hermann Rorschach.

The initial tool came under severe attack in the 1950s and 1960s, in part because it lacked standardized procedures and a set of norms (averaged results from the general population).

Standardization is important because seemingly trivial differences in the way an instrument is administered can affect a person’s responses to it. Norms provide a reference point for determining when someone’s responses fall outside an acceptable range.

In the 1970s John E. Exner, Jr., then at Long Island University, ostensibly corrected those problems in the early Rorschach test by introducing what he called the Comprehensive System. This set of instructions established detailed rules for delivering the inkblot exam and for interpreting the responses, and it provided norms for children and adults.

In spite of the Comprehensive System’s current popularity, it generally falls short on two crucial criteria that were also problematic for the original Rorschach: scoring reliability and validity. A tool possessing scoring reliability yields similar results regardless of who grades and tabulates the responses. A valid technique measures what it aims to measure: its results are consistent with those produced by other trustworthy instruments or are able to predict behavior, or both.

To understand the Rorschach’s scoring reliability defects, it helps to know something about how reactions to the inkblots are interpreted. First, a psychologist rates the collected reactions on more than 100 characteristics, or variables. The evaluator, for instance, records whether the person looked at whole blots or just parts, notes whether the detected images were unusual or typical of most test takers, and indicates which aspects of the

inky swirls (such as form or color) most determined what the respondent reported seeing.

Then he or she compiles the findings into a psychological profile of the individual. As part of that interpretive process, psychologists might conclude that focusing on minor details (such as stray splotches) in the blots, instead of on whole images, signals obsessiveness in a patient and that seeing things in the white spaces within the larger blots, instead of in the inked areas, reveals a negative, contrary streak.

For the scoring of any variable to be considered highly reliable, two different assessors should be very likely to produce similar ratings when examining any given person's responses. Recent investigations demonstrate, however, that strong agreement is achieved for only about half the characteristics examined by those who score Rorschach responses; evaluators might well come up with quite different ratings for the remaining variables.

Equally troubling, analyses of the Rorschach's validity indicate that it is poorly equipped to identify most psychiatric conditions—with the notable exceptions of schizophrenia and other disturbances marked by disordered thoughts, such as bipolar disorder (manic-depression). Despite claims by some Rorschach proponents, the method does not consistently detect depression, anxiety disorders or psychopathic personality (a condition characterized by dishonesty, callousness and lack of guilt).

Moreover, although psychologists frequently administer the Rorschach to assess propensities toward violence, impulsiveness and criminal behavior, most research suggests it is not valid for these purposes either. Similarly, no compelling evidence supports its use for detecting sexual abuse in children.

Other problems have surfaced as well. Some evidence suggests that the Rorschach norms meant to distinguish mental health from mental illness are unrepresentative of the U.S. population and mistakenly make many adults and children seem maladjusted. For instance, in a 1999 study of 123 adult volunteers at a California blood bank, one in six had scores supposedly indicative of schizophrenia.

The inkblot results may be even more misleading for minorities. Several investigations have shown that scores for African-Americans, Native Americans, Native Alaskans, Hispanics, and Central and South Americans differ markedly from the norms. Together the collected research raises serious doubts about the use of the Rorschach inkblots in the psychotherapy office and in the courtroom.

### Doubts about TAT

ANOTHER PROJECTIVE TOOL—the Thematic Apperception Test (TAT)—may be as problematic as the Rorschach. This method asks respondents to formulate a story based on ambiguous scenes in drawings on cards. Among the 31 cards available to psychologists are ones depicting a boy contemplating a violin, a distraught woman clutching an open door, and the man and woman who were mentioned at the start of this article. One card, the epitome of ambiguity, is totally blank.

The TAT has been called “a clinician's delight and a statistician's nightmare,” in part because its administration is usually not standardized: different clinicians present different numbers and selections of cards to respondents. Also, most clinicians interpret people's stories intuitively instead of following a well-tested scoring procedure. Indeed, a recent survey of nearly 100 North



## RORSCHACH TEST Wasted Ink?

*“It looks like two dinosaurs with huge heads and tiny bodies. They're moving away from each other but looking back. The black blob in the middle reminds me of a spaceship.”*

Once deemed an “x-ray of the mind,” the Rorschach inkblot test remains the most famous—and infamous—projective psychological technique. An examiner hands 10 symmetrical inkblots one at a time in a set order to a viewer, who says what each blot resembles. Five blots contain color; five are black and gray. Respondents can rotate the images. The one above is an inverted version of an Andy Warhol rendering; the actual Rorschach blots cannot be published.

Responses to the inkblots purportedly reveal aspects of a person's personality and mental health. Advocates believe, for instance, that references to moving animals—such as the dinosaurs mentioned above—often indicate impulsiveness, whereas allusions to a blot's “blackness”—as in the spaceship—often indicate depression.

Swiss psychiatrist Hermann Rorschach probably got the idea of showing inkblots from a European parlor game. The test debuted in 1921 and reached high status by 1945. But a critical backlash began taking shape in the 1950s, as researchers found that psychologists often interpreted the same responses differently and that particular responses did not correlate well with specific mental illnesses or personality traits.

Today the Comprehensive System, meant to remedy those weaknesses, is widely used to score and interpret Rorschach responses. But it has been criticized on similar grounds. Moreover, several recent findings indicate that the Comprehensive System incorrectly labels many normal respondents as pathological.

### THE AUTHORS

SCOTT O. LILIENFELD, JAMES M. WOOD and HOWARD N. GARB all conduct research on psychological assessment tools and recently collaborated on an extensive review of research into projective instruments that was published by the American Psychological Society [see “More to Explore,” on page 87]. Lilienfeld and Wood are associate professors in the departments of psychology at Emory University and the University of Texas at El Paso, respectively. Garb is a clinical psychologist at the Pittsburgh Veterans Administration Health Care System and the University of Pittsburgh and author of the book *Studying the Clinician: Judgment Research and Psychological Assessment*.

American psychologists practicing in juvenile and family courts discovered that only 3 percent relied on a standardized TAT scoring system. Unfortunately, some evidence suggests that clinicians who interpret the TAT in an intuitive way are likely to overdiagnose psychological disturbance.

Many standardized scoring systems are available for the TAT, but some of the more popular ones display weak

“test-retest” reliability: they tend to yield inconsistent scores from one picture-viewing session to the next. Their validity is frequently questionable as well; studies that find positive results are often contradicted by other investigations. For example, several scoring systems have proved unable to differentiate normal individuals from those who are psychotic or depressed.

A few standardized scoring systems

for the TAT do appear to do a good job of discerning certain aspects of personality—notably the need to achieve and a person’s perceptions of others (a property called “object relations”). But many times individuals who display a high need to achieve do not score well on measures of actual achievement, so the ability of that variable to predict a person’s behavior may be limited. These scoring systems currently lack norms and so are

## THEMATIC APPERCEPTION TEST

# Picture Imperfect

The Thematic Apperception Test (TAT), created by Harvard University psychiatrist Henry A. Murray and his student Christiana Morgan in the 1930s, is among the most commonly used projective measures. Examiners present individuals with a subset (typically five to 12) of 31 cards displaying pictures of ambiguous situations, mostly featuring people. Respondents then construct a story about each picture, describing the events that are occurring, what led up to them, what the characters are thinking and feeling, and what will happen later. Many variations of the TAT are in use, such as the Children’s Apperception Test, featuring animals interacting in ambiguous situations, and the Blacky Test, featuring the adventures of a black dog and its family.

Psychologists have several ways of interpreting responses to the TAT. One promising approach—developed by Boston University psychologist Drew Westen—relies on a specific scoring system to assess people’s perceptions of others (“object relations”). According to that approach, if someone wove a story about an older woman plotting against a younger person in response to the image visible in the photograph at the right, the story would imply that the respondent tends to see malevolence in others—but only if similar themes turned up in stories told about other cards.

Surveys show, however, that most practitioners do not use systematic scoring systems to interpret TAT stories, relying instead on their intuitions. Unfortunately, research indicates that such “impressionistic” interpretations of the TAT are of doubtful validity and may make the TAT a projective exercise for both examiner and examinee.



not yet ready for application outside of research settings, but they merit further investigation.

### Faults in the Figures

IN CONTRAST TO THE RORSCHACH and the TAT, which elicit reactions to existing images, a third projective approach asks the people being evaluated to draw the pictures. A number of these instruments, such as the frequently applied Draw-a-Person Test, have examinees depict a human being; others have them draw houses or trees as well. Clinicians commonly interpret the sketches by relating specific “signs”—such as features of the body or clothing—to facets of personality or to particular psychological disorders. They might associate large eyes with paranoia, long ties with sexual aggression, missing facial features with depression, and so on.

As is true of the other methods, the research on drawing instruments gives reason for serious concern. In some studies, raters agree well on scoring, yet in others the agreement is poor. What is worse, no strong evidence supports the validity of the sign approach to interpretation; in other words, clinicians apparently have no grounds for linking specific signs to particular personality traits or psychiatric diagnoses. Nor is there consistent evidence that signs purportedly linked to child sexual abuse (such as tongues or genitalia) actually reveal a history of molestation. The only positive result found repeatedly is that, as a group, people who draw human figures poorly have somewhat elevated rates of psychological disorders. On the other hand, studies show that clinicians are likely to attribute mental illness to many normal individuals who lack artistic ability.

Certain proponents argue that sign approaches can be valid in the hands of seasoned experts. Yet one group of researchers reported that experts who administered the Draw-a-Person Test were

## OTHER PROJECTIVE TOOLS

# What's the Score?

### Hand Test

Subjects say what hands pictured in various positions might be doing. This method is used to assess aggression, anxiety and other personality traits, but it has not been well studied.

### Handwriting Analysis (Graphology)

Interpreters rely on specific “signs” in a person’s handwriting to assess personality characteristics. Though useless, the method is still used to screen prospective employees.

### Lüscher Color Test

People rank colored cards in order of preference to reveal personality traits. Most studies find the technique to lack merit.

### Play with Anatomically Correct Dolls

Research finds that sexually abused children often play with the dolls’ genitalia; however, that behavior is not diagnostic, because many nonabused children do the same thing.

### Rosenzweig Picture Frustration Study

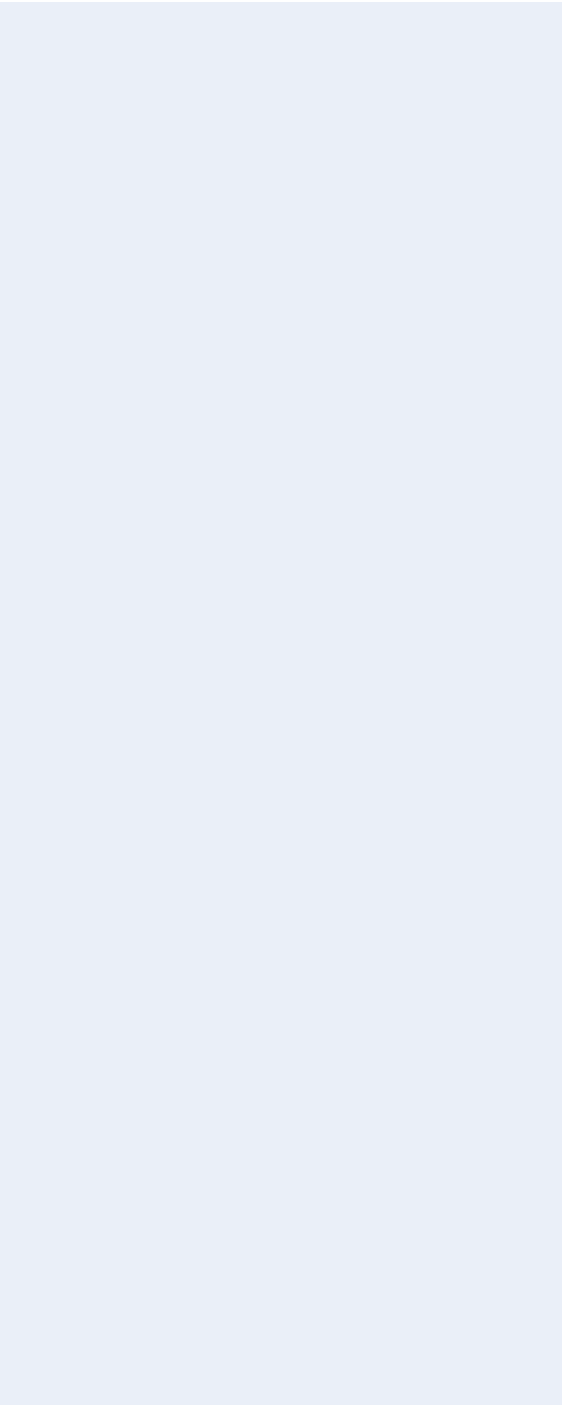
After one cartoon character makes a provocative remark to another, a viewer decides how the second character should respond. This instrument, featured in the movie *A Clockwork Orange*, successfully predicts aggression in children.

### Sentence Completion Test

Test takers finish a sentence, such as, “If only I could . . .” Most versions are poorly studied, but one developed by Jane Loevinger of Washington University is valid for measuring aspects of ego development, such as morality and empathy.

### Szondi Test

From photographs of patients with various psychiatric disorders, viewers select the ones they like most and least. This technique assumes that the selections reveal something about the choosers’ needs, but research has discredited it.



less accurate than graduate students at distinguishing psychological normality from abnormality.

A few global scoring systems, which are not based on signs, might be useful. Instead of assuming a one-to-one correspondence between a feature of a drawing and a personality trait, psychologists who apply such methods combine many aspects of the pictures to come up with a general impression of a person's adjustment. In a study of 52 children, a global

scoring approach helped to distinguish normal individuals from those with mood or anxiety disorders. In another report, global interpretation correctly differentiated 54 normal children and adolescents from those who were aggressive or extremely disobedient. The global approach may work better than the sign approach because the act of aggregating information can cancel out "noise" from variables that provide misleading or incomplete information.

Our literature review, then, indicates that, as usually administered, the Rorschach, TAT and human figure drawings are useful only in very limited circumstances. The same is true for many other projective techniques, some of which are described in the box on the preceding page.

We have also found that even when the methods assess what they claim to measure, they tend to lack what psychologists call "incremental validity": they rarely add much to information that can


be obtained in other, more practical ways, such as by conducting interviews or administering objective personality tests. (Objective tests seek answers to relatively clear-cut questions, such as, “I frequently have thoughts of hurting myself—true or false?”) This shortcoming of projective tools makes the costs in money and time hard to justify.

### What to Do?

SOME MENTAL HEALTH professionals disagree with our conclusions. They argue that projective tools have a long history of constructive use and, when administered and interpreted properly, can cut through the veneer of respondents’ self-reports to provide a picture of the deepest recesses of the mind. Critics have also asserted that we have emphasized negative findings to the exclusion of positive ones.

Yet we remain confident in our conclusions. In fact, as negative as our overall findings are, they may paint an overly rosy picture of projective techniques because of the so-called file drawer effect. As is well known, scientific journals are more likely to publish reports demonstrating that some procedure works than reports finding failure. Consequently, researchers often quietly file away their negative data, which may never again see the light of day.

We find it troubling that psychologists commonly administer projective instruments in situations for which their value has not been well established by multiple studies; too many people can suffer if erroneous diagnostic judgments influence therapy plans, custody rulings or criminal court decisions. Based on our findings, we strongly urge psychologists to curtail their use of most projective techniques and, when they do select such instruments, to limit themselves to scoring and interpreting the small number of variables that have been proved trustworthy.

Our results also offer a broader lesson for practicing clinicians, psychology students and the public at large: even seasoned professionals can be fooled by their intuitions and their faith in tools that lack strong evidence of effectiveness. When a substantial body of research demonstrates that old intuitions are wrong, it is time to adopt new ways of thinking. 

## HOW OFTEN THE TOOLS ARE USED Popularity Poll

In 1995 a survey asked 412 randomly selected clinical psychologists in the American Psychological Association how often they used various projective and non-projective assessment tools, including those listed below. Projective instruments present people with ambiguous pictures, words or things; the other measures are less open-ended. The number of clinicians who use projective methods might have declined slightly since 1995, but these techniques remain widely used.

PROJECTIVE TECHNIQUES	USE ALWAYS OR FREQUENTLY	USE AT LEAST OCCASIONALLY
<i>Rorschach</i>	43%	82%
<i>Human Figure Drawings</i>	39%	80%
<i>Thematic Apperception Test (TAT)</i>	34%	82%
<i>Sentence Completion Tests</i>	34%	84%
<i>CAT (Children’s version of the TAT)</i>	6%	42%
NONPROJECTIVE TECHNIQUES*	USE ALWAYS OR FREQUENTLY	USE AT LEAST OCCASIONALLY
<i>Weshler Adult Intelligence Scale (WAIS)</i>	59%	93%
<i>Minnesota Multiphasic Personality Inventory-2 (MMPI-2)</i>	58%	85%
<i>Weschler Intelligence Scale for Children (WISC)</i>	42%	69%
<i>Beck Depression Inventory</i>	21%	71%

\* Those listed are the most commonly used nonprojective tests for assessing adult IQ (WAIS), personality (MMPI-2), childhood IQ (WISC) and depression (Beck Depression Inventory).

SOURCE: “Contemporary Practice of Psychological Assessment by Clinical Psychologists,” by C. E. Watkins et al. in *Professional Psychology: Research and Practice*, Vol. 26, No. 1, pages 54–60; 1995.

### MORE TO EXPLORE

The Rorschach: A Comprehensive System, Vol. 1: Basic Foundations. Third edition. John E. Exner. John Wiley & Sons, 1993.

The Comprehensive System for the Rorschach: A Critical Examination. James M. Wood, M. Teresa Nezworski and William J. Stejskal in *Psychological Science*, Vol. 7, No. 1, pages 3–10; January 1996.

Studying the Clinician: Judgment Research and Psychological Assessment. Howard N. Garb. American Psychological Association, 1998.

Evocative Images: The Thematic Apperception Test and the Art of Projection. Edited by Lon Gieser and Morris I. Stein. American Psychological Association, 1999.

Projective Measures of Personality and Psychopathology: How Well Do They Work? Scott O. Lilienfeld in *Skeptical Inquirer*, Vol. 23, No. 5, pages 32–39; September/October 1999.

The Scientific Status of Projective Techniques. Scott O. Lilienfeld, James M. Wood and Howard N. Garb in *Psychological Science in the Public Interest*, Vol. 1, No. 2, pages 27–66; November 2000. Available at [www.psychologicalscience.org/newsresearch/publications/journals/pspi1\\_2.html](http://www.psychologicalscience.org/newsresearch/publications/journals/pspi1_2.html)