THE HOLTZMAN INKBLOT TECHNIQUE:  
A REVIEW  
KENNETH R. GAMBLE  
Loyola University of Chicago

After a brief description of the rationale, construction, and structure of the Holtzman Inkblot Technique, this study critically reviews 10 years (1959–1969) of research with this instrument. An impressive body of positive findings is in evidence, but many more studies are needed in most areas covered by this review. The various technical refinements and extensions of this instrument that have been produced should facilitate the appearance of much needed future research. The dearth of relevant studies renders an evaluative comparison with the Rorschach impossible at this time. Implications of certain issues raised by R. Schafer in 1948, M. D. Ainsworth in 1954, and R. Holt in 1968 in relation to Rorschach validity studies are discussed in reference to research with the Holtzman technique.

Although Holtzman recently provided summaries of research on the Holtzman Inkblot Technique (Holtzman, 1966, 1968), his presentations were not evaluative, not much information was given concerning the designs of the studies reviewed, and a number of studies were not included. The present review (a) presents a brief description of the rationale, construction, and structure of this instrument; (b) provides an evaluative review of the literature from 1959 to 1969, inclusive; and (c) discusses the general character of this research in light of previous criticisms of Rorschach validity studies. It should be noted that this review attempts to be primarily informative and secondarily evaluative. It is hoped that the inclusion of more than the usual amount of descriptive material with respect to specific studies will facilitate the future appearance of much needed additional research that is indicated in several areas covered by this review.

The authors of the Holtzman Inkblot Technique (HIT) have described their goal as one of developing “a new inkblot technique having scores of demonstrated psychometric value while still preserving the rich, qualitative essence of the Rorschach [Holtzman, Thorpe, Swartz, & Herron, 1961, p. 7].” It is possible to think of the HIT as a natural successor to the important position occupied by its well-known predecessor, the Rorschach. This is not to suggest that deposition of the Rorschach is imminent or even highly likely at this time. Considering, however, that the HIT is another inkblot technique constructed along lines established in the Rorschach tradition, and further, that the newer instrument claims psychometric superiority over the Rorschach, the eventual replacement of the Rorschach by the HIT seems to be more than a remote logical possibility. In view of this and the fact that the Rorschach is probably the most frequently used psychological test in clinics and hospitals, a review of research conducted with the HIT would seem to be of some importance at this point.

To fully appreciate the significance of the HIT, it is necessary to keep in mind the well-documented deficiencies of the instrument that it is most likely to replace, the Rorschach. Probably in their most exaggerated form, these deficiencies have been presented by Zubin (1954) and quoted in Holtzman et al. (1961) as follows:

(1) failure to provide an objective scoring system free of arbitrary conventions and showing high inter-scorer agreement; (2) lack of satisfactory internal consistency or test-retest reliability; (3) failure to provide cogent evidence for clinical validity; (4) failure of the individual Rorschach scoring categories to relate to diagnosis; (5) lack of prognostic or predictive validity with respect to outcome of treatment or later behavior; (6) inability to differ-

1 The author would like to thank Frank Kohler for his encouragement and for his helpful criticisms of an earlier version of this manuscript.

Requests for reprints should be sent to Kenneth R. Gamble, 1545 West Pratt Boulevard, Chicago, Illinois 60626.
entiate between groups of normal subjects; and (7) failure to find any significant relationships between Rorschach scores and intelligence or creative ability [p. 5].

Without the qualification that this list represents an exaggerated view, one would find it difficult to reconcile the fact that the Rorschach is the most frequently used instrument in clinical practice today with the assumption that at least a majority of clinicians would be reality oriented enough to recognize its deficiencies and abandon such a faulty instrument! Perhaps some of the cognitive dissonance generated by the consideration of these facts and assumptions can be reduced if we recognize with Holt (1968), for example, that maybe most of the validity studies reported have missed the whole point of Rorschach usage and the interpretations it generates. Maybe there is no effective means of quantifying the clinical insight that the "master testers" seem to have. Still, and without rejecting the values implied in Holt's analysis of research with projective techniques, one can welcome an effort to improve psychometrically a method (such as the Rorschach) even if this endeavor may lead, in the final analysis, to a new method bearing little resemblance to the "wisdom" from which it sprang.

Structure and Standardization of the HIT

The HIT consists of two "parallel" sets, each containing 45 inkblots that were selected and matched on the basis of item analyses of hundreds of experimental blots. An important consideration in the selection of a given blot was its ability to elicit small detail, space, color, and shading determined responses from a subject. Attention to these variables explicitly places the HIT squarely within the tradition begun by Rorschach, in which it is assumed that percepts organized on the basis of different determinants may be taken to reflect inter- and intraindividual variations in personality organization. Another reason for selecting blots with strong "pulling power" with respect to the several locations and determinants relates to one alleged point of psychometric superiority of the HIT over the Rorschach: the requirement of the HIT that only one response be given to each card. The latter provision was made to counteract the confounding of response frequency with the frequency and variety of various locations and determinants in a subject's record. Since only one response is allowed on the HIT, and, since experience has shown that form-determined wholes are the usual first responses to inkblots, special attention to "pulling power" with respect to other inkblot variables may be viewed as a necessary compensatory measure.

The final selection of a blot for inclusion in the final two forms of the HIT (A and B) was based on three empirical criteria: (a) Its ability to discriminate between a group of college student volunteers and a group of hospitalized psychotics; (b) the amount that a blot contributed to the total scores on several variables (such as location, color, shading, movement) in each sample; and (c) inter- and intrascorer reliability. The results of the initial standardization study were most encouraging. Realizing the gross (or just simply inadequate) nature of the distinction between college students and hospitalized psychotics, it is still impressive that, on all variables measured, the groups were clearly differentiated and in a manner that is highly consistent with previous Rorschach findings: Scores for the college group were significantly higher on Location, Form Appropriateness, Form Definiteness, Movement, and Shading. No significant differences were found on mean Color scores, but greater variance was observed in the psychotic group. Split-half reliability coefficients were found to range from a low of .31 to a high of .96 with most falling in the 70s and 80s (Holtzman et al., 1961).

As Holtzman described them, the blots that were finally selected cover a wide range of stimulus variation, giving the individual ample opportunity to reveal certain aspects of his mental processes and personality by projecting his thoughts onto otherwise meaningless inkblots. Twelve of the inkblots in Form A are black or gray, two are monochromatic, eleven are black with a bright color also present, and the remaining twenty are multicolored. Most of the blots have rich shading variations which help to elicit texture responses. A similar distribution of color, shading, and form qualities is present in Form B [Holtzman, 1968, p. 139].
Except for the marked asymmetry of many HIT blots, it is interesting to note the essential similarity between the stimulus properties of the HIT and those of the traditional Rorschach Blots. (Is this new bottles for old wine?)

Variables Derived from HIT Responses

The following 22 variables comprise the scoring system developed for the HIT (Holtzman, 1968):

Reaction Time (RT)—the time, in seconds, from presentation of the inkblot to the beginning of the primary response.

Rejection (R)—score 1 when the subject returns the inkblot to the examiner without giving a scorable response.

Location (L)—tendency to break down the inkblot into smaller fragments; score 0 for use of the whole blot, 1 for use of a large area of the blot, 2 for use of smaller areas of the blot.

Space (S)—score 1 for responses involving a figure-ground reversal where white space constitutes the figure and the inkblot is the ground.

Form Definiteness (FD)—a 5-point scale ranging from a score of 0 for a concept having completely indefinite form (“squashed bug”) to a score of 4 for highly specific form (“man on horse”).

Form Appropriateness (FA)—goodness of fit of the form of the concept to the form of the inkblot; score 0 for poor, 1 for fair, and 2 for good form.

Color (C)—importance of both chromatic and achromatic color as a determinant; score 0 when not used, 1 when used only in a secondary manner (as in the Rorschach Form Primary [FC]), 2 when color is a primary determinant but some indefinite form is present or implied (as in the Rorschach Color Primary [CF]), and 3 when color is primary and no form is present (as in the Rorschach Response [C]).

Shading (Sh)—importance of shading or texture as a determinant; score 0 when not used, 1 when used only in a secondary manner, and 2 when shading is a primary determinant.

Movement (M)—a 5-point scale for measuring the degree of movement, tension, or dynamic energy projected into the percept by the subject, regardless of content; score 0 for none, 1 for static potential (sitting, looking, resting), 2 for casual movement (walking, talking), 3 for dynamic movement (dancing, weeping), and 4 for violent movement (whirling, exploding).

Pathognomic Verbalization (V)—a 5-point scale ranging from 0 (no pathology present) to 4 (very bizarre verbalizations) for measuring the degree of disordered thinking represented by fabulations, fabulized combinations, queer responses, incoherence, autistic logic, contaminations, self-references, deteriorated responses, and absurd responses.

Integration (I)—score 1 when two or more adequately perceived blot elements are organized into a larger whole.

Human (H)—score 0 for no human content present, 1 for parts of human beings, featureless wholes, or cartoon characters, 2 for differentiated humans or the human face if elaborated.

Animal (A)—score 0 for no animal content, 1 for animal parts, and 2 for whole animals.

Anatomy (At)—score 0 for no penetration of the body wall, 1 for X rays, medical drawings, or bone structures, and 2 for viscera or soft internal organs.

Sex (Sx)—score 0 for no direct sex references, 1 for socially accepted sexual activity and expressions (“buttocks,” “kissing”), and 2 for blatant sex references (“penis”).

Abstract (Ab)—score 0 if no abstract concept is present, 1 if abstract elements are secondary, and 2 if the response is wholly abstract, for instance, “Reminds me of happiness.”

Anxiety (Ax)—a 3-point scale for rating the degree of anxiety apparent in the content of the response as reflected in feelings or attitudes (“frightened animal”), expressive behavior (“girl escaping”), symbolic responses (“dead person”), or cultural stereotypes of fear (“witch”); score 1 when debatable or indirect, and score 2 when clearly evident.

Hostility (Hs)—a 4-point scale for rating degree of hostility apparent in the content of the response, with increasing score as hostility moves from vague or symbolic expressions to more direct, violent ones in which human beings are involved.
Barrier (Br)—score 1 for reference to any protective covering, membrane, shell, or skin that might be symbolically related to the perception of body-image boundaries.

Penetration (Pn)—score 1 for concepts symbolic of body penetration.

Balance (B)—score 1 where the subject expresses concern for the symmetry-asymmetry dimension of the inkblot.

Popular (P)—score 1 if a popular response is given, popular responses being defined statistically for specific areas of the inkblots in earlier normative studies of the HIT.

The decision to include a variable in the HIT scoring scheme was based on several considerations (Holtzman, 1966; Holtzman et al., 1961): (a) Could most of the traditional Rorschach scores be derived from the new system? (b) Is it theoretically possible to have any score from 0 to 45? (c) Can high scoring agreement be reached among trained individuals? (d) Is the variable relevant to the study of personality? and (e) Are the variables logically (but not necessarily empirically) independent of one another? It seems that all of these conditions have been met. Rorschach Whole (W), Determinants (D), or Common Detail (d) would be Location 0, 1, or 2, respectively. Rorschach Human Movement (M) is Movement scored 2 or higher. Rorschach Entire Human (H) is Human 1 or 2. Rorschach FC, CF, or C would correspond to the number of chromatic cards coded 1, 2, or 3, respectively, for Color. Item b above is evident from the structure of the scoring scheme, and Item e is a matter of data (discussed below). Item d needs to be decided on the basis of validity research, and Item e seems to be a matter of judgment.

Reliability

To be sure, the characteristic of the HIT that stands out most is its improvement, reliability-wise, over other inkblot techniques. Though reliability estimates are shown to vary markedly from sample to sample (Kobler & Doiron, 1968), variations can usually be attributed to skewness, a restriction in the range of variation in a sample, or chance. And, in general, the magnitude of the coefficients is quite acceptable. Interscorer correlations range from .89 to .995 when highly trained scorers are used to evaluate individual protocols. Even examiners whose training is limited to little more than a reading of the examples in the scoring guide produce interscorer correlations ranging from .73 to .89. Intrascorer consistency ranges from .89 to .97. Test-retest stability has been shown to range from .39 (for P) to .82 (for L) over a period of one week. Test-retest correlations obtained when the HIT was administered at an interval of one year ranged from .24 (P) to .75 (L) (Holtzman et al., 1961). Holtzman believes that the average stability coefficient observed over several studies is high enough so as to provide stable measures of personality phenomena, but low enough to indicate that the HIT is sensitive to normal variations in personality over time (Holtzman et al., 1961).

Technical Refinements and Extensions

Gorham (1967) reported on the reliability and validity of computer-scored (Moseley, Gorham, & Hill, 1963), group-administered HIT protocols on a group of 145 college students. The 17 HIT variables studied were L, R, FD, C, Sh, M, I, H, A, At, Sx, Ab, Ax, Hs, Br, Pn, and P. The criteria for computer-score validation were an expert scorer’s values obtained when protocols were scored in the usual manner. The correlation between the average of three hand scorers and the computer scores either equaled or approached the interscorer reliability of the hand scorers. Overall, the correlations between hand scores and computer scores ranged from a low of .54 (Ab) to a high of 1.00 (R). An additional validity index reported is the fact that a factor analysis of computer scores resulted in eight factors that were nearly identical to factors obtained when hand-scored protocols were analyzed in this way.

Norms exist for the computer-scored HIT variables L, R, FD, C, Sh, M, I, H, A, At, Sx, Ab, Hs, Br, Pn, P (Gorham, Moseley, & Holtzman, 1968) on over 5,000 subjects including high school and college students and United States Navy enlistees; clinical subjects include Veterans Administration and state hospital depressives, schizophrenics, psychoneurotics, alcoholics, and chronic brain-syndrome patients; cultural samples include university students from Argentina, Australia,
Colombia, Denmark, Germany, Mexico, Hong Kong, Hungary, India, Japan, Lebanon, Nigeria, Panama, Turkey, Venezuela, and Yugoslavia. The usefulness of these norms as applied to individually administered and scored inkblots should be evident from the finding that for the various methods, correlations have either equaled or approached the magnitude of interscorer reliability (Gorham, 1967; Holtzman, Moseley, Reinehr, & Abbott, 1963).

In addition, Herron (1963) has shown that a short form (first 30 items) of the group-administered HIT produces means and standard deviations for most variables that are highly similar to those observed on the 45-item test. Using a criterion setting internal stability at .70, then, the variables R, L, FD, FA, C, V, and H can be used with confidence to detect group differences using the short form.

The availability of these technical extensions should increase the attractiveness of the HIT “package,” especially for research involving group comparisons. It can probably be assumed that the time required to administer and score the Rorschach has deterred many researchers from using this instrument. The high degree of concurrent validity shown in the computer-scored, group-administered, and short form of the HIT should facilitate the inclusion of inkblot variables in studies where they might not appear under other circumstances. Psychological researchers are, it has been said, only human, and this alone may predispose them to seek out relatively quick and easy measures for research. (The Taylor Manifest Anxiety Scale is probably legendary in this respect.)

A technical refinement of a different sort is implied in a study by Megargee (1966a). It will be recalled that one of the deficiencies of the Rorschach that the HIT sought to overcome was the contaminating influence that variable response productivity has on other inkblot scores. The solution provided by the HIT involved limiting the subject to one response per card so as to rule out the influence of this factor at least on a single card. It is questionable, however, that this procedure has solved this problem. Megargee (1966a) pointed out that though response frequency is controlled in the HIT (a supposed psychometric advantage over the Rorschach), there is still the possibility that productivity (of which response frequency is an index) may not be controlled as response length is still free to vary. Response length (RL) refers to the mean number of words used in each scorable response. Megargee (1966a) correlated RL with all of the HIT variables in two samples (84 college students and 75 male juvenile delinquents) and found several highly significant positive relationships in both samples between RL and M, Ab, Ax, Hs, and Br. The pervasive influence of RL was shown when this variable was factor analyzed along with the other HIT variables and found to load (.71) on a most important factor of the HIT (Factor I, defined by M, I, B, H, and P and accounting for more variance than any other factor).

A second study (Megargee, 1966a) sought to determine more precisely the relationship between RL and M, an important index of personality functioning. A group of high-M subjects and a group of low-M subjects (on the basis of HIT, Form B) were further subdivided into groups encouraged to give long responses (30 words) and groups encouraged to give short responses (10 words), on Form A. These groups were then assessed for amount of M. Manipulating RL in this way had a highly significant impact ($p < .001$) on M production. The low-M-short group produced a mean M of 12.33 on the second administration, high-M-long produced 71.20, high-M-short produced 25.53, low-M-long produced 43.67. These effects are clear-cut and strong. Megargee (1966a) pointed out that RL may be an important variable entering into many relationships between inkblot perception and personality, including intelligence, experimental manipulations, and examiner effects. Needless to say, future investigations using the HIT must be regarded as inadequate unless they come to terms with the “problem” of RL. In another vein, it is conceivable, in view of Megargee’s (1966a) experimental findings, that RL will turn out to be an important variable in its own right.

**Direct Comparisons with the Rorschach**

The initial comparison between the HIT and the Rorschach is reported in Holtzman
et al. (1961). In this study, HIT and Rorschach protocols (scored by Beck’s system) of 72 high school students were compared on eight variables judged to be comparable in the two systems. When Rorschach scores were adjusted for response frequency, the correlations between these two instruments were found to range from .30 (FA) to .79 (A), and all correlations were significant beyond the .01 level. Extensive multivariate analyses of these data were carried out by Bock, Haggard, Holtzman, Beck, and Beck (1963). A canonical regression analysis (judged by these investigators as most relevant to the question of Rorschach-HIT equivalence) revealed that 60% of the stable canonical variation in the Rorschach scores was predictable by 14 HIT scores (8 HIT scores had to be eliminated because of a high number of zero frequencies). Thus, while it is apparent that the Rorschach and HIT have much in common, it is also evident that these two techniques are far from measuring the same variables as well, and it is questionable whether or not the same dimensions on both tests ought to be given the same interpretations.

A somewhat unusual comparative study by Otten and Van de Castle (1963) compared Form A of the HIT with the Rorschach on a semantic differential, using bipolar adjectives such as clean–dirty, active–passive. The 10 Rorschach cards were mixed in with the HIT cards, and the subjects responded to each card on the differential. The HIT cards were found to cover all of the semantic space covered by the Rorschach, and some patterns emerged on the HIT that were not present on the Rorschach. It is interesting to speculate that the latter find is related to the effort to put more “pulling power” in the HIT cards to compensate for the one-response requirement.

Whitaker (1965) correlated Pathognomic Verbalization (V) scores on these two tests administered to 45 psychiatric inpatients. The HIT method of scoring V was applied to the Rorschach protocols as well. Independent scorers were enlisted to analyze a subset (N = 19) of these protocols. The correlation obtained between Scorer A’s Rorschach with Scorer B’s HIT scores was .76 ($p < .01$).

Scorer A’s Rorschach and HIT scores for the entire sample of 45 patients correlated .94. The latter coefficient is undoubtedly somewhat inflated due to the operation of criterion contamination as only one scorer was involved. On the other hand, the correlation of .76 may be somewhat attenuated because of examiner unreliability due to criterion disagreement. In either case, Rorschach and HIT V appear to have much in common.

In view of the effort to maintain some of the Rorschach tradition in the construction of the HIT, there are surprisingly few studies effecting direct comparisons between the two techniques. The studies that have been carried out are all positive in one sense or another, but it is clear that, even aside from the psychometric differences, the instruments are not that similar. Actually, the kind of comparison that is needed—one in which diagnoses rather than scores are compared—has not as yet been reported in the literature. If the correlations between the test scores were higher than they are, this would obviously be unnecessary. But since the scores do not correlate that perfectly, a diagnostic comparative study appears essential. Besides the direct comparison of diagnoses, other interesting comparisons could be made as well. For example, what of the dimensions that these two techniques do not have in common? Do they have additional diagnostic significance? Or, if the diagnostic formulations turn out to be much the same for the two techniques, is there more stability in the HIT diagnosis on repeated tests? In general, a great deal of comparative work that could have been done with these instruments has not been done.

**Examiner and Set Influences**

Experimental evidence from several sources agrees that “examiner” variables have a marked effect on certain HIT variables.

Hamilton and Robertson (1966) randomly assigned 90 college students to “warm,” “cold,” and “neutral” conditions of examination, the same examiner serving under all three conditions. Significant differences were found in scores on FD, FA, M, I, H, and Word Productivity. The “warm” condition resulted in the most productive protocols, and the “cold” condition resulted in the least pro-
ductive protocols with respect to these variables.

Simkins (1960) reported a study designed to test the influence of the presence and degree of examiner reinforcement on HIT responses. Subjects were matched as closely as possible with respect to similarity of content, location, and determinants given on an initial administration of the HIT. Three matched groups were obtained in this way and given one of the following treatments during a second testing session: strong reinforcement for a given response category (“very good”), weak reinforcement (“um-hmm”), or no reinforcement. Scores termed by Simkins as D scores were computed for each subject based on the change in the number of responses in a reinforced category from the first to the last session. Another difference score, D', was based on the difference between scores for matched pairs under strong- and weak-reinforcement conditions. Weak reinforcement was found to be more effective in both the location and content categories, while strong reinforcement was most effective with determinant scores. There was no significant difference, however, between weak and strong reinforcement. In comparison to matched controls, the content and determinant dimensions showed the largest reinforcement effects, and location was most resistant to reinforcement-induced change.

Marwit and Marcia (1967) studied the effect of “experimenter bias” on the number of responses (not usually free to vary) given by college students to five selected HIT cards. The examiners were 36 volunteer students from an undergraduate course in experimental psychology. Thirteen of an own-hypothesis group expected a high number of responses, and 6 expected a low number. Twelve examiners in a given-hypothesis group were told to expect a high number of responses, and 5 were told to expect a low number. Both the own and given conditions effected the responsivity of the subjects significantly (p < .01). Subjects whose examiners had high expectations gave an average of approximately 18 responses, while the low-expectancy examiners elicited an average of around 11 responses. Moreover, when mean responses per card were plotted as a function of card sequenve, an increasing function “learning” curve appeared with significant linear (p < .01) and quadratic (p < .05) trends in the high-expectancy groups. This finding suggests that a learning process (in the experimenter or subject, or both!) may be mediating examiner effects.

A somewhat different kind of “examiner” influence was studied by Herron (1964). This investigator administered the HIT to college students (a) under the standard instructional set and (b) as a test of intellectual ability. Under the second condition, significant decreases were found in Pn, Hs, A, and V and a slight (but not highly significant) increase in FA. These results were interpreted as reflecting a slight “tightening-up” of cognitive process under the intelligence test set.

To be sure, the findings of studies dealing with examiner influences are interesting in their own right. Still, they may be criticized for their failure to address themselves to broader issues in the assessment of personality through inkblot perception. Knowledge that examiner and set variables influence specific scores is certainly useful information. These findings remind us of the necessity of strict standardization of the test situation where other influences are our main concern. It must be remembered, however, that these studies reflect experimentally induced as opposed to natural examiner differences, and generalizations from the former to the latter cannot be made without qualifications. Further, the assertion that diagnoses or personality descriptions are affected by examiner variables does not follow from the fact that individual scores are so affected. To demonstrate that diagnoses, for instance, are a function of examiner influences requires studies in which diagnoses rather than scores are the dependent variable.

**Studies of External Validity**

**Developmental Changes**

Organismic developmental theory (Werner, 1948) views the development of living systems as proceeding in the direction of increasing differentiation of part functions and processes, followed by increasing hierarchic integration achieved by the subordination of these part functions. As applied to inkblot
perception, significant developmental trends may be predicted when responses are defined in terms of these principles.

Thorpe and Swartz (1965) reported a developmental study of HIT responses that used the developmental principles outlined above. Five criterion age groups from 5 to 22 years of age were isolated, and care was taken to insure that each group contained equal numbers of males and females ($N = 586$). All 10 HIT variables studied showed significant age-group differences (most at $p < .001$), but there were no significant sex differences nor were there any significant Sex $\times$ Age interactions. Six variables—$FA$, $FD$, $I$, $M$, $H$, and $Sh$—showed steadily increasing means with increasing age. The changes observed in $FA$, $FD$, and $I$ are in keeping with Werner's notion that cognitive development proceeds in a direction away from loosely organized and amorphous percepts. The increases in $H$ and $M$ probably reflect an increase in the integrative capacities of the organism in the area of perceptual cognitive development. The systematic increase in $Sh$ responses probably reflects an increasing sensitivity to very subtle aspects of the stimuli and an ability to integrate these with form quality to produce richer percepts. Increases in $L$ followed by a decline in $L$ in the oldest group follows the trend of increasing differentiation followed by integrative efforts ($W$ responses) resulting in the actual lowering of $L$ in the most advanced group—theoretically a sign of hierarchic integrative efforts. Animal responses went up and down in seesaw fashion in the groups studied. $V$ first declined and then rose again, as did $C$. These latter trends are more difficult to interpret developmentally, and attempts at theorizing should be contingent upon their replication in other studies.

In a follow-up study (Thorpe & Swartz, 1966) that involved a partial replication of a previous study (Thorpe & Swartz, 1965), the HIT was administered individually to normal subjects who were 6.7, 9.7, and 12.7 years of age. On the basis of the developmental trends found earlier, eight HIT variables were selected for study: $FA$, $FD$, $I$, $M$, $H$, $C$, $Sh$, and $V$. Significant age-group differences were found again by analysis of variance for all variables except $Sh$. The only sex difference to emerge was on $H$, where females had slightly higher scores than males. In line with the earlier study, $FA$, $FD$, $I$, $M$, and $H$ showed increasing mean scores across the age groups employed. At variance with the earlier data, $V$ was found to decrease with age. As in the earlier study, $L$ showed a decline from 6.7 to 9.7 years of age and then a slight increase at age 12.7.

Swartz, Lara Tapia, and Thorpe (1967) provided intercultural validation for the developmental trends observed in the studies described above. In this study, 300 normal Mexican school children living in Mexico City were divided into criterion-age groupings of 6.7, 9.7, and 12.7 years. Of the 11 HIT variables studied, $RT$, $FD$, $FA$, $C$, $M$, $I$, and $H$ showed significant $(p < .001)$ age-group differences. No Sex $\times$ Age interactions were found, but females at each age level were found to have a higher mean $L$ score ($p < .01$) and a lower $M$ score ($p < .01$) than males. All of the variables showing significant age effects with the exception of $C$ showed steadily increasing means as a function of age. Five variables in this study—$FA$, $FD$, $M$, $I$, and $H$—showed developmental trends entirely consistent with those observed in other samples drawn from the United States. One trend, not observed in other samples, was the linear increase in $RT$ in the Mexican sample. This may turn out to be an important cultural deviation that future research ought to clarify. In the main, this study provided strong support for the developmental significance of trends found in common with earlier studies.

Sanders, Holtzman, and Swartz (1968) carried out an excellent longitudinal study of developmental trends in the HIT $C$ variable. On the basis of previous results and theory, $C$ responses ("3") in the HIT system) should predominate in early childhood, $CF$ (HIT "2") in later childhood, and $FC$ (HIT "1") in adolescence and adulthood. The 323 subjects in the study were divided into three age groups: 6.7, 9.7, and 12.7 years. These groups were observed over a 10-year period (from 5.7 to 15.7 years of age). As predicted, significant declines in $C$ were observed as a function of age. Significant monotonic increments in $CF$ and $FC$ with age were not observed, however.
Taken together, these studies provide strong support for the view that certain HIT scores provide reliable indexes of developmental changes in cognitive organization. The nature of these changes and the particular variables reflecting this change strongly support Werner's (1948) notion that cognitive development proceeds along lines of increasing differentiation and integration, but certain details of the trends may require additional assumptions. The decline and resurgence of $V$ found in one study, for example, may reflect the capacity to integrate somewhat pathological elements in basically normal personalities, or it may reflect stresses of an emotional nature brought on by puberty. Ignoring slight ambiguities in the interpretation of minor aspects of the data, the consistency of findings with respect to theoretically relevant and developmentally meaningful conceptions is noteworthy. A logical next step for research in this area to take would involve testing subjects in middle and old age to determine whether the HIT is sensitive to the reverse trends of decreasing differentiation and hierarchic integration predicted by organismic developmental theory.

Cross-Cultural Studies

Knudsen, Gorham, and Moseley (1966) tested the universality of the HIT variable $P$ by group administration of the HIT to subjects in Mexico, Germany, Hong Kong, Denmark, and the United States. The criterion for designating a given response as a popular was that it be present in at least one out of every seven protocols. Universality was found for 13 out of Holtzman's 25 original $P$ across all five cultures. Eight universal $P$s were found which were very close to the remaining 12. On 15 inkblots, the core universal concept was found to be Person. Other concepts found to be universal were animal, fowl, face, mask, person-riding-animal, landscape, butterfly, moth, fish, seahorse, and rain or storm.

Derogatis, Gorham, and Moseley (1968) sought to determine the relationship between the structural (physical) ambiguity and the interpretive ambiguity of the HIT, the latter being defined in terms of the variability of responses made to the stimulus. Another goal of this study was to determine the amount of agreement on these dimensions among individuals with diverse backgrounds (psychologists and American, Mexican, Chinese, and German students). Structural ambiguity was assessed by ratings on a 5-point scale (high to low), and interpretive ambiguity was measured by the total number of different words given in response to each HIT card by each of four samples of 100 subjects from Mexico, Germany, China, and the United States.

Group administration was used (Swartz & Holtzman, 1963), and scoring of the protocols was carried out by means of the computer program developed by Gorham (1967). The structural ambiguity ratings of the four student samples correlated significantly on all cards (.70-.80), and the correlation between the psychologists' ratings and the average pooled student ratings was .90. The extent of agreement on interpretive ambiguity is reflected in the correlations among the various national samples. These relationships were positive and only moderately high (.44-.55) but all significant. The average correlation between structural and interpretive ambiguity was $- .35 \ (p < .01)$. As the authors pointed out, these findings suggest that structural ambiguity is a concept that may be culture free, as indicated by the between- and within-cultural consistencies observed. The findings with respect to interpretive ambiguity are not as strong, but they tend to point in the direction of a fair amount of intercultural agreement.

These cross-cultural studies strongly support the intercultural validity of certain processes involved in inkblot perception. Other studies are finding interesting intercultural differences in HIT performance as well. Swartz (1966) found that 6-12-year-old Mexicans obtain lower scores on most HIT variables. Tamm's (1966) results were nearly identical to these. When she compared both American and Mexican bilingual first, fourth, and seventh graders in a Mexico City school, she found consistent and stable differences across all age levels. The differences that emerged in these studies are interesting, and they should be replicated. In view of the high degree of intercultural commonality observed in HIT responses, any consistently observed differences are bound to deserve extra
attention (for a comprehensive account of longitudinal and cross-cultural research with the HIT currently in progress, see Holtzman, Diaz-Guerrero, Swartz, & Lara Tapia, 1968).

Cognitive Processes

Data obtained from the HIT standardization sample of seventh-grade children showed significant correlations from .20 to .31 between measures of general mental ability and R (inverse), L, M, FA, Sh, I, Ax, Hs, and Br. Thorpe and Swartz (1963) replicated the positive relationship between mental ability and R in low (mean of 12 rejections) and high (mean of 6 rejections) groupings of seventh-grade children. Similarly, Holtzman, Gorham, and Moran (1964) found significant positive correlations between vocabulary and I, M, and FA in chronic, paranoid schizophrenics (details of this study are presented below), and Holtzman (1968) reported an average correlation of .27 between vocabulary levels of school children and M, FD, I, H, Hs, Br, and P. Whether or not these relationships hold up in cognitively more stable normal or neurotic adults remains to be determined—and ought to be.

Young (1959) intercorrelated 10 cognitive personality measures, some of which were measures of the field dependence–independence construct. Two measures—a “coping” score and an “introspective” score—were derived from the HIT. For males, low introspectiveness was associated with low-analytic ability (r = -.29) but not with field dependence, and low-coping ability was associated with field dependence (.32 and .44—two tests). For females, low introspectiveness was associated with field dependence (.30), and low-coping ability was associated with field dependence (.43) and with low-analytic ability (.36). The manner in which these results were reported made it impossible to determine exactly what HIT variables were involved in these relationships.

In a study comparing three groups of college students differing only in verbal quantitative discrepancies in ability, Sanders, Mef- ferd, and Brown (1960) found no significant differences on HIT variables, with the exception of RT. Subjects with consonant verbal quantitative abilities were found to have the longest RT, followed in order by low-verbal–high-quantitative subjects and high-verbal–low-quantitative subjects. Subjective impressions of the examiners, however, correctly placed 60% of the subjects in the three groups based only on the HIT responses, but this mode of data analysis was not elaborated by the authors.

In a study of convergent and divergent thinking in talented adolescents, Clark, Veldman, and Thorpe (1965) included 16 HIT variables in an analysis of several measures related to these cognitive styles. There were no significant main HIT effects associated with convergent thinking, but L, M, Ax, Hs (p < .01), C, and Pn (p < .05) successfully discriminated between high and low groups in divergent thinking. High-divergent thinkers used larger blot areas (low L) and produced higher scores on all of the other five variables mentioned. These results were interpreted as indicating that divergent thinkers are capable of giving freer rein to their imaginations when given the opportunity. In addition, however, these same subjects were more responsive to the stimulus characteristics of the inkblots (low L, high C), so that their imaginativeness was not achieved at the sacrifice of reality contact. This pattern is very reminiscent of typical findings in Rorschach studies of creativity.

Richter and Winter (1966) found significance differences between high- and low-creative potential (Myers-Briggs Type Indicator) female college students (matched for age and verbal ability) on FD (p < .01), C (p < .001), M (p < .0005), H (p < .005), I (p < .05), V (p < .025), Ax (p < .0005), Hs (p < .025), and Ab (p < .05). The high-creative potential group achieved higher scores than the low-creative group on all of these variables. Interestingly, L (reflecting whole or detail responses) did not differentiate these groups. One reason offered by the authors to explain the latter finding is that the HIT does not differentiate between complex, accurately perceived wholes reflecting creative ability and whole responses that do not require much in the way of the superior integrative capacities that highly creative persons are said to possess.

Using Megargee’s (1966a) finding that re-
Response length (RL) is a significant mediating (or confounding) variable, Gray (1969) tried to determine whether HIT measures of primary process (an important component of creative functioning according to many psychoanalytic-oriented writers) would still correlate significantly with measures of creativity when RL was partialed out. Holt's (1963) system was used to operationally define primary process, and six tests of divergent thinking comprised a creativity battery. The correlation between these measures was .23 ($p < .05$). When productivity scores (RL on the HIT and number of responses on the creativity tests) were partialed out, the correlation between primary process and creativity fell to .06. Thus, the correlations that have been observed in the past between these variables were very likely due to a third variable, productivity. And if this is true with respect to the HIT studies, it is probably doubly true in the many studies using the Rorschach where variation not only in RL but also in number of responses has probably combined to produce many spurious relationships between inkblot indexes and creativity abilities. This finding creates difficulties for psychoanalytic theories of creative functioning and also (more creatively) presents a new challenge to future investigations of this dimension.

Body Image

Cleveland and Fisher (1960) attempted to replicate and extend earlier findings concerning the relationships between body image and physical symptoms choice. These studies had suggested that rheumatoid arthritics, as a group, may be characterized in the following terms: They have (a) strong conflicts centered on the containment of hostile feelings, (b) during some period of their lives diverted these impulses into vigorous muscular activity through competitive sports and rugged outdoor activity, (c) a body image that emphasizes the containing and protecting aspects of the body boundary, and (d) strong unconscious exhibitionistic desires.

Certain features of the body images of arthritics (and those of persons suffering other stress-type ailments such as ulcers) have been formalized in a body-image scale wherein responses referring to objects with well-defined boundaries ("turtle with a hard shell" for example) are counted and summarized in a $Br$ score. Arthritics have typically been found to score relatively high on this scale. By contrast, ulcer patients, who tend to project their body exterior as weak and vulnerable, score low on the $Br$ scale but relatively high on a $Pn$ scale (scored for responses like "bleeding wound").

Though earlier studies of body image utilized the $Br$ and $Pn$ scales in conjunction with the Rorschach test, more recent investigations have utilized the HIT. Cleveland and Fisher's data (1960), using the HIT, upheld earlier investigations with the Rorschach by finding significantly higher $Br$ scores in arthritics as opposed to ulcer patients ($p < .001$), while the reverse trend was found on $Pn$ scores ($p < .05-.10$). A hostility score derived from the HIT responses failed to differentiate these groups, but in comparison with norms reported on normal subjects, both of these groups were high in projected hostility. Incidentally, consistent with theoretical formulations concerning arthritics, this group was found to have a significantly higher tendency to engage in vigorous competitive sports ($p < .001$) and unusually strong interests in cooking activities ($p < .001$) (all subjects, by the way, were male Veterans Administration inpatients).

Cleveland and Morton (1962) attempted to extend the body-image construct to the area of social interaction. HIT $Br$ scores were obtained from 69 Veterans Administration, open ward patients, and these were related to indexes of group behavior derived from a sociometric questionnaire. The high-$Br$ subjects were found to receive significantly more descriptions as influential, active, independent, and goal striving, and they were less influenced by group opinion than low-$Br$ subjects. Low-$Br$ subjects were described in contrasting terms as more frequently selfish and parochial. These findings are in keeping with theoretical conceptions relating $Br$ scores to body boundary definiteness.

Renik and Fisher (1968) were successful in experimentally inducing boundary changes in normal males. HIT $Br$ scores assessed on a test-retest basis served as the criteria in
this study. In between the first and second HIT administrations, subjects were asked either (a) to direct their attention to their skin and muscles, (b) to focus on the interiors of their bodies, or (c) not to focus on the body. Exterior-focusing males did not show a significant increase in Br (unlike females), but interior-focusing males did show a significant decrease in Br (unlike females). When the data for females (Fisher & Renik, 1966) and males (Renik & Fisher, 1968) were combined, the exterior-focus condition was found to significantly increase Br when compared to the interior-focus condition, but not as compared to the control condition. The interior-focus condition did produce a significant Br decrement when compared to the control conditions. In comparing the results of these two studies, it is interesting to note that females became “exterior” more easily, while males became “interior” more easily. The explanation offered for this sex difference is that females are more practiced in the alteration of the body surface (use of makeup, adornments, etc.) than are males. When men are found who have a tendency to ornament their body surface by tattoo (Moshier, Oliver, & Dolgan, 1967), they are found to have significantly higher Br scores on the HIT than men who are not so inclined. In both cases (females and tattooed males), increased Br seems to reflect a tendency toward boundary-strengthening preoccupations. (No significant sexual deviations were found in the men.)

A single dissenting set of findings (Hartley, 1967) found no significant relationships between HIT Br scores and site of symptoms (internal versus external) in a sample of college students. Caution must be exercised in interpreting these results, however, as evidently none of the reported symptoms were serious enough to require hospitalization, and 45 of the 83 subjects studied could not endorse symptoms in “often” or “serious enough to require medical treatment” categories.

Taken together, these studies reinforce the significance of the body-boundary construct in personality research, and they tend to support the construct validity of the HIT Br and Pn scores. Other studies relevant to these HIT dimensions are presented elsewhere in this review.

Empathy

Mueller and Abeles (1964) utilized the HIT to test the relationship between the production of human movement responses and the capacity for empathy in advanced students in clinical and counseling psychology. The HIT responses were scored for M, H, and FA. Several components of empathy were assessed. The findings showed that perceived or projected movement is significantly related to only one component of empathy, the accuracy with which others perceive the subject’s behavior. In view of this result, the “high M” person was felt to be a person who makes more information about himself available for appraisal by others—an interesting but rather unusual notion of “empathy.” It would seem that this study could have benefited considerably by the inclusion of nonclinical control groups (e.g., experimental students, graduate physicists), as one possible explanation for the generally weak findings may be the restriction in the range of variation of the empathy variable by using clinical students on the one hand, and the restriction in range of H, M, and FA that must have occurred because of the intellectual level of this sample.

Fernold and Linden (1966) tested the hypothesis that the HIT variable H is positively related to social isolation and functional pathology (schizophrenia). Contrary to the advice of Holtzman et al. (1961), human, human detail, and humanlike responses received equal credit in assigning an H score. Using a group sociometric technique for assessing social isolation and empathy, no relationship was found between these variables and H. When social interest scores on the Strong Vocational Interest Test were related to H, a significant relationship was found and in the predicted direction. To test the relationship between H and psychopathology, a group of twenty-one normal 23–55-year-old firemen were compared to a group of twenty 21–57-year-old chronically hospitalized male schizophrenics. The number of H responses given by these two groups differed significantly (p < .01) and in the predicted direction.

M and H are indeed important variables in Rorschach lore, and it would be important to
have information on the validity of these as assessed by the HIT. The two studies presented here can only be taken as suggestive for reasons of design already discussed. Another criticism to be made of these studies—indeed of many validation studies of both the HIT and the Rorschach—is that they are poorly conceived. As Ainsworth (1954) pointed out “there are no unqualified interpretative hypotheses attached to these values as such [p. 417].” Yet researchers continue to design studies as if M, H, or any other determinant is supposed to reveal some psychological predisposition. There is nothing “wrong” with these studies per se—as long as their results are not interpreted to reflect upon the validity of Rorschach usage as it is actually carried out in clinical practice. This criticism can be made of most of the studies reviewed herein.

**Aggression**

In comparisons of delinquent with nondelinquent adolescents and extremely delinquent with less delinquent subjects, Megargee (1965b) set out to test the relationship between HIT Br scores and aggressive behavior. The delinquent sample in this study was found to have a mean Br score significantly lower than the nondelinquents represented in Holtzman et al.’s (1961) samples (p < .001). As a control on RL (known to be significantly related to Br), the delinquent sample was divided into less delinquent and extremely delinquent subsamples. Though there was no significant difference in verbosity (i.e., RL) between these groups, the extremely delinquent subgroup was found to be significantly lower in mean Br score than the less delinquent group (p < .001). Further confirmation of the inverse relationship between overt aggression and Br scores was obtained by correlating the latter with aggressive ratings made by counselors who were in daily contact with the delinquents. The correlation between these indexes was .23 (p < .05) and in the direction predicted (a low score on the aggressiveness rating scale means high aggression). These results were interpreted as supportive of the notion that Br scores may be an index of both body boundary and ego identity. The sample of delinquents studied here must represent a doubly extreme case, as the normal adolescents they were compared with themselves had lower Br scores than most other nonclinical groups (Holtzman et al., 1961).

In a study primarily directed toward expanding the normative data on the HIT, Megargee (1965a) compared HIT protocols of 75 male delinquents with those of non-delinquent adolescents reported by Holtzman et al. (1961). The latter group comprised a sample of seventh and eleventh graders. According to age criteria alone, the delinquent group would be expected to fall midway between these groups. Instead, they were significantly lower than the seventh graders on most variables and identical to or lower than the seventh graders, the only exceptions to this being R and At scores on which the delinquents were significantly higher than the seventh graders and slightly higher than the seventh graders (the .01 level of significance was maintained throughout). One way of characterizing these results is to describe the delinquent sample as grossly immature in its makeup, and this corresponds favorably to the salient features of delinquent behavior usually reported.

In another normative study, Megargee (1966b) compared the HIT protocols of white and Negro male juvenile delinquents who were matched for mental age. Three HIT variables significantly differentiated whites from Negroes: V (whites higher, p < .03), At (Negroes higher, p < .04), and P (whites higher, p < .02). Three other variables showed noteworthy trends, but statistical significance was not achieved (whites were higher on RT, FA, and C). Interestingly, these groups did not perform differently on the Thematic Apperception Test (TAT) or the Rosenzweig Picture-Frustration Study.

In another study, Megargee (1966c) dichotomized a group of male juvenile delinquents into a moderately assaultive and an extremely assaultive group and compared the subgroups on a variety of measures including the HIT. The general hypothesis of this study was that the extremely assaultive group would be lower on measures of aggressiveness and higher on measures of control than the moderately assaultive group and groups of
nonassaultive delinquents selected for control purposes (incorrigibles, property offenders). The main findings of this study with respect to the HIT were as follows: (a) Comparison groups did not differ significantly on a $Hs$ scale derived from the inkblot responses; (b) the prediction that the extremely assaultive group would be highest on a $M - C$ index (presumably a measure of overcontrol) was tested by simply subtracting each subject's $C$ score from his $M$ score. This hypothesis was supported as the extremely assaultive group was significantly higher on $M - C$ than all three of the other comparison groups ($p < .061$) and higher than the moderately assaultive group in particular ($p < .059$). As an additional check on the overcontrol hypothesis, pure $C$ responses were tabulated for all groups, and again the extremely assaultive group differed significantly from the moderately assaultive groups ($p < .045$), the former producing fewer pure $C$ responses. (The difference between extremely assaultive and the other groups combined was significant only at $p < .111$, but all differences were in the predicted direction—showing relative overcontrol in the extremely assaultive group.) If for no other reason than the fact that an attempt was made to conceptualize the relationship between the HIT variables and the predicted behavior in some psychologically meaningful way, this study is to be applauded.

Megargee and Cook (1967) correlated a number of HIT-derived aggression indexes as well as TAT-derived aggression scales to observer and self-report ratings of aggression in 76 juvenile delinquents. Complex patterns of relationships were found depending on the scale and the overt criterion used for measuring aggression. In general, the TAT scales were found to relate most closely to school conduct (preoffense behavior), while the HIT-derived scales related more closely to physical (not verbal) aggression assessed after arrest during interviews and by means of direct observation. The Holtzman et al. (1961) hostility scale correlated $- .24$ ($p < .05$) with "total physical aggression" and $.25$ ($p < .05$) with a global rating of aggression (low scores indicating greater aggressiveness) but $.03$ with verbal aggression. One must concur with the authors' conclusion: "For the clinician who might wish to use these scales in the prediction of overt aggression in the individual case, the results are quite discouraging [Holtzman et al., 1961, p. 58]." It is necessary to add, however, that this study provided no information on nondelinquents' aggression for purposes of comparison. It is possible, too, that the restriction of range in aggression due to such a select sample may have attenuated important relationships in this behavior domain. A design that ought to be attempted in this area is one in which unselected normals who are both high and low on aggression (inkblot assessed) are further subdivided into manifest aggressive and controlled groups for exploratory study on the factors that mediate between fantasy and overt aggression.

**Diagnosis**

Many of the studies included in this section are not, strictly speaking, studies of differential diagnosis in the sense of diagnostic efficiency. Considering the latter type of diagnostic study, there is a surprising dearth of information concerning the diagnostic validity of the HIT. Several studies, though not directly concerned with diagnostic efficiency, provide data bearing on important aspects of psychiatric classification, so they are presented in this section.

**Schizophrenia.** In order to determine the major dimensions underlying schizophrenic thought processes, Holtzman et al. (1964) factor analyzed the intercorrelations among HIT scores, taken from several perceptual cognitive tests, age, education, and length of illness. The subjects were 99 chronic, paranoid schizophrenic men who had been hospitalized for from 20 months to 22 years. Of the eight factors needed to account for the common variance, only those that were found to have substantial loadings on the HIT variables are presented here. A factor called Integrated Ideation was defined primarily by $M$, $I$, $Br$, and $R$ (negatively). Stimulus Sensitivity was defined positively by $C$, $Sk$, and $B$, and negatively by $FD$. Length of illness had a loading of $- .40$ on this factor, which points to a relationship between chronicity and sensitivity. Pathological Verbalization was defined by $Ax$, $HS$, and $V$. Though Conceptual Autism had its highest loading on an object-sorting task
revealing an “open-private” orientation (abstract sortings having private or idiosyncratic meanings), the HIT variables Ab, At, Sx, Fa (negatively), Affect Arousal (AA) (since discarded from the HIT), and Pn also defined this factor. Sexual Concern was found to be a small but sharply delineated dimension in this sample, and it was defined primarily by Sx, AA, and S with some loading on L and Pn as well. The authors pointed out that this factor has not appeared in analyses of data from normal populations. In general the results of this study were consistent with other studies of schizophrenic thinking that have used divergent measures. In the interest of questions of the concurrent factorial validity of the HIT, it should be emphasized that common variance shared by both the HIT variables and the other measures was such as to reinforce the pathological significance of very high or very low scores on several of the HIT variables.

As part of the standardization of the HIT (Holtzman et al., 1961), data were obtained on chronic paranoid schizophrenics, depressed neurotics and psychotics, and mentally retarded subjects. As compared to normal reference groups, the chronic schizophrenics were found to obtain significantly higher scores on R, V, Pn, At, Sx, and significantly lower scores on L, FD, Fa, Sh, M, I, Br, and P.

Utilizing 16 HIT scores, Moseley (1963) performed a discriminant function analysis and developed a formula for the classification of schizophrenics, depressives, and normals. When applied to schizophrenics versus normals, classification was 88% correct. When a formula was applied to schizophrenic versus depressives, correct classification was achieved in 78% of the cases, and 71% correct classification resulted for depressives versus normals.

Cleveland (1960) made repeated tests on a group of hospitalized schizophrenics (N = 25) to determine the relationship between changes in body image during therapy and personality reorganization as assessed by a morbidity measure and a psychiatrist’s disposition of the patients at the end of the study. The HIT was administered prior to chemotherapy and again after 5 and 13 weeks (ABA and BAB sequence of the alternate forms was used). All protocols were scored for Br and Pn. A significant rho of .60 was found between morbidity ranking and Pn for predrug versus fifth-week testing, and a rho of .61 was obtained for predrug versus thirteenth-week testing. Br scores yielded no significant correlations in this analysis. By the second criterion (case disposition), the same trend is observed and reflected in significantly greater decrements in Pn scores in discharged patients as compared to patients retained in the hospital. In an unusual way, these findings tend to support other diagnostic studies involving schizophrenics. The changes in Pn were interpreted as reflecting “firming up and defining” of body image boundaries as a function of therapeutic improvement. Presumably, this represents an improvement in that aspect of schizophrenic behavior which has been described as “the inner experience of personal dissolution at the periphery . . . [Szasz, 1957, p. 127].” If this is so, it is puzzling that increases in Br were not found to be associated with clinical improvement.

As part of a study of empathy (described elsewhere in this review), Fernold and Linden (1966) found that hospitalized male schizophrenics gave significantly fewer H responses than a comparison group of normal males (p < .01).

Process-reactive distinction. Becker (1956) suggested that process and reactive schizophrenics may be differentiated on the basis of cognitive developmental principles (Werner, 1948). Based on the salient features of the process syndrome that relate to the presence of a relatively undifferentiated personality structure (as opposed to a differentiated and hierarchically differentiated personality, see Werner, 1948), Becker (1956) predicted that process schizophrenics would show more regressive and immature cognitive processes on the Rorschach than would reactive schizophrenics. Using the conception that development proceeds in the direction of increasing differentiation and integration (Werner, 1948), a genetic-level scoring system was developed for Rorschach responses. At the extreme undifferentiated end of the scale, for example, are Amorphous and Minus Whole responses, as well as Confabulations and
Perseverations. At the highly differentiated end of the scale are Whole and Detail responses of high form level. Intermediate to these are Vague Whole responses and mediocre Detail responses. With the Elgin Prognostic Scale as a measure of the process-reactive dimension, Becker's (1956) hypothesis was supported. Separate analyses were made of male and female schizophrenics and the Elgin scale, and the Rorschach genetic-level scores were found to correlate $- .599 (p < .01)$ for the men and $- .679 (p < .001)$ for the women. More recently, Steffy and Becker (1961) replicated this finding using the HIT. The correlation between Elgin scores and genetic level derived from the HIT was $- .36 (p < .05)$ on a sample of 36 hospitalized schizophrenics. When duration of hospitalization was partialed out to control for lower Elgin ratings and improved inkblot performance (both a function of length of hospitalization), the correlation rose to $- .46$. The authors concluded that the HIT shows promise of producing good measures of degree of pathology in schizophrenia.

Using a different measure of the process-reactive continuum, Ullman and Eck (1963) extended these findings to male schizophrenics who were either discharged from the hospital or leaving on a trial visit ($N = 48$). The HIT variables used in this study were $V$, $I$, and $FA$. When $V$, $I$, and $FA$ were added together to form an inkblot summary score, the correlation between the process-reactive measure and inkblot summary score was found to be $+.47 (p < .001)$. (The correlation in this study is in the positive direction because of the scoring procedures used.) These findings corroborate those found earlier, but it should be noted that this sample was atypical, and the relationship between the HIT variables and the genetic-level hypothesis was not discussed, although (based on the results of developmental studies discussed above) the use of Integration and Form Appropriateness would indicate that some developmentally relevant dimensions were being tapped by the inkblot summary score measure.

Degree of autism. Using a somewhat complex design, Hill (1966) studied the degree of affect aroused by chromatic HIT cards as a function of stimulus strength ("brilliance" rated by 26 hospital staff members) and degree of autism ("high" and "low" as assessed by selected scales of the Sixteen Personality Factor Questionnaire). The dependent variable in this study, "affect arousal," was operationalized in terms of ratings of nonverbal behavior, Holtzman's discarded $AA$ scale, and $FD$. Analysis of variance of the $FD$ scores found stimulus strength and degree of autism to be significant variables. More specifically, $FD$ to strong colors was lower than $FD$ to weak colors, and $FD$ was lower for high-autistic subjects than for low-autistic subjects. On low stimulus-strength cards, the high- and low-autistic subjects did not differ. With respect to $AA$ scores, strong color cards produced greater affect arousal than weak cards, and high-autistic subjects scored higher than low-autistic subjects in $AA$. Also, degree of autism interacted significantly with stimulus strength ($p < .05$) indicating that low-to-high shifts in stimulus strength had more impact on high-autistic subjects than on low-autistic subjects. The trends in the overt behavior ratings were identical to those with respect to $AA$ scores, except that there was no difference between the groups under conditions of low stimulus strength of color. The effects observed in all three dependent variables support the conclusion that stimulus strength and degree of autism and the interactions between these variables are important determiners of color responses on the HIT. The interaction effects may also be interpreted to mean that some HIT cards are more discriminating with respect to degree of autism than others.

Organicity. Only one study reported the use of the HIT to differentiate between normal and brain-damaged individuals (Barnes, 1964). It is claimed (Holtzman, 1968) that Barnes could discriminate these groups with 80% accuracy. The difficulty with this contention is the fact that many other measures were used besides the HIT in a multivariate design by this investigator, and the manner in which the study was reported did not allow one to assess the precise contribution of the HIT to differentiation or whether the discriminating variables were defined a priori or post hoc. More details about this study are needed before final evaluation can be made, however. Anxiety and neuroticism. Megargee and
Swartz (1968) correlated scores on the Extraversion (E) and Neuroticism (N) scales of the Maudsley Personality Inventory (MPI) with the 21 scores of the HIT. The 89 subjects (40 women and 49 men) in the study were undergraduates of the University of Texas, and group administration procedures were used for both the HIT (Swartz & Holtzman, 1963) and the MPI. The resulting 19 scores (four HIT variables were dropped because of extreme skewness) were subjected to a principal-axis analysis. Contrary to predictions, no significant correlations were found between HIT C, M, or H and Extraversion. Thus it appears that extraversion-introversion does not mean the same thing when assessed by the HIT and the MPI. (The authors claim that their prediction was made on the basis of “inkblot lore.”) It is interesting to note, however, that Klopfer, Ainsworth, Klopfer, and Holt (1954) insist that introversion-extraversion is a different concept from inkblot derived introversion-extravasion. In view of this confirmation of the difference between these constructs, it would be interesting to determine how they differ in other respects. There were several significant correlations between MPI Neuroticism and HIT variables: R (−.24), FA (−.23), M (.23), V (.23), Ax (.31), and Hs (.24). Though none of these correlations are very high, they are significant and all in the expected direction considering inkblot hypotheses of neurotism. The factor analysis carried out in this study to determine whether any of the HIT factors could be interpreted as Extraversion-Introversion or Neuroticism, as Eysenck (1965) suggested, showed that Extraversion is relatively independent of the HIT as its loadings on any factors even slightly reminiscent of extraversion were quite low (.44 and .41). The loadings of the Neuroticism scale were similarly low (the highest being .49) but appropriately on a factor labeled disordered thought processes and emotional disturbance.

Van de Castle and Spicher (1964) used the HIT to investigate the phenomenon of color disturbance. High- and low-scoring groups of college students on two tests of anxiety and a test of neuroticism were administered 8 chromatic cards and 15 mixed cards. Subjective disturbance was measured by means of a 16 pair–item semantic differential. No evidence for color shock was found. In fact, both groups rated the chromatic cards in more favorable terms than the nonchromatic or mixed cards. The authors suggested that the term “novelty shock” be substituted for “color shock” as disruption in “set” seemed to mediate disruption in test performance, and this could possibly come about were an achromatic card suddenly to appear after several chromatic cards.

For evidence of the construct validity of several HIT variables, a study by Herron (1965) provided an interesting diversion from the usual approach. Herron studied the relationship between eyelid conditionability and certain HIT factor marker variables presumed to measure anxiety and neurotism. Conditionability was dichotomized with failures defined as less than 15% conditioned responses in 60 trials. Using Holtzman et al.’s (1961) version of HIT factors, the Neuroticism factor (4x, V) was found to be positively related to conditionability. V correlated .30 and Ax .36 (p < .05) with conditionability. Holtzman’s Cognitive factor was also positively related to conditionability (r with I = .44; r with H = .22). It is interesting to note that one meaning of this cognitive factor involves the notion that a person with high scores (I) can ignore irrelevant aspects of a stimulus situation in the interest of organizing a “good” percept. Similarly, in eyelid conditioning studies, it has been found that the elimination of distracting stimuli significantly speeds up the rate of conditioned response acquisition (Porter, Engel, Brady, & Kropp, 1964). Another interesting trend in these data was shown by the positive correlation between the Sensitivity factor (C = .22 and Sh = .26) and conditionability.

Swartz and Swartz (1968) related the Test Anxiety Scale for Children to 11 HIT variables using a sample of 120 normal children (60 high anxiety and 60 low anxiety) within 2 weeks of ages 6.7, 9.7, and 12.7 years. The four variables found to differ significantly between the groups were M (low-anxiety subjects giving more), Pn (high-anxiety group giving more), and AA (high-anxiety group giving more). Two of these differences—M and AA—replicate the findings of an earlier
study of anxiety in children (Swartz, 1965). The findings with respect to $M$ diverge sharply from previous studies of anxiety and Rorschach responses (Sarason, Davidson, Lighthall, & Waite, 1958).

It is significant that these studies and others (Barger & Sechrest, 1961; Doris, Sarason, & Berkowitz, 1963; Holtzman et al., 1961, p. 180) generally fail to show strong relationships between questionnaire- and rating-assessed anxiety and HIT indexes. These and other findings have prompted Holtzman et al. (1961) to conclude:

Anxiety and Hostility as scored in the Holtzman Inkblot Technique are strictly ratings at a fantasy level which are not necessarily related in any simple, direct way to overt behavior that is judged to be anxious or hostile [pp. 180-181].

Further, we would add that it is a task for theory to try to conceptualize just in what way—if not “simple” and “direct”—fantasy productions should relate to self-report and overt behavior.

Alcoholism. Cleveland and Sikes (1966) compared 70 hospitalized alcoholics with 50 nonalcoholics on HIT $Br$, $Pn$, Decadence (any response involving deterioration), and Water responses. Chi-square analysis of the data indicated that while the criterion groups did not differ on $Br$ scores, alcoholics were significantly higher on $Pn$ ($p < .02$), Decadence ($p < .001$), and Water responses ($p < .001$). The results on these and other variables generally supported the hypothesis that alcoholics differ from nonalcoholics in the perception of their bodies as “dirty, disgusting, and in a state of decay” and in the diffuseness of their body boundary concepts. It is important to call attention to the stability of their perceptions: no change was found when these subjects were tested again after 90 days of therapy. With respect to the finding that alcoholics gave more water responses, the only thing that can be said is that here is one more positive finding in a body of literature where both positive and negative results seem equally likely. One more thing should be said about the structure of this study in relation to diagnostic differentiation. The chi-square analyses showed that there were more alcoholics giving high numbers of responses in the categories of $Pn$, Decadence, and Water responses as compared to low frequencies tabulated, but the frequencies in these categories for nonalcoholics were considerable, indicating that, had a different design been used, a great deal of overlap may have been shown between these groups.

Mayfield (1968) hypothesized that alterations in inkblot perception would occur under acute alcohol intoxication because of the alterations in personality that are assumed to result under such conditions. On the basis of previous results and theorizing, this researcher predicted an increase in $C$ and a decrease in $FA$ in an acute alcohol injection condition, as compared to a dextrose (placebo) injection condition. The subjects were paid, white, male volunteers ($N = 12$) from 26 to 50 years of age. Both Forms A and B of the HIT were used in a before-after design. No significant changes occurred on any HIT variable with the exception of $C$, which increased from 3.3 to 4.8 in the alcohol condition and decreased from 3.9 to 2.2 in the placebo-control condition. These results were viewed as in line with typical psychodynamic theorizing on the effects of intoxication (release of inhibitions, impulsivity).

Miscellaneous diagnostic studies. Moseley, Duffey, and Sherman (1963) factor analyzed the HIT variables along with scores on an Inpatient Multidimensional Psychiatric Scale (IMPS) and the Minnesota Multiphasic Personality Inventory (MMPI). The subjects were 82 Veterans Administration patients who had been diagnosed as neurotic depressive (46) or psychotic depressive (36), the latter category including involutionals (2), manic depressives (5), psychotic depressives (7), and schizo-affectives (22). Ten factors emerged in this study. Three factors were defined only by HIT variables: Organized Normal Ideation (highest loadings on $I$, $M$, and $H$), Scoring Factor ($A$, $L$, and negatively by $R$), and Fantasized Hostility with Anxiety ($Hs$, $Ax$, and $M$). The HIT was found to have little relationship to the MMPI scales, but significant relationships were observed in the pattern of loadings when the clinical ratings of overt behavior (IMPS) were considered. A Withdrawal and Disorientation factor showed significant loadings on IMPS Disorien-
tation and Grandiose Expansiveness and HIT Sx, and V. A Fluctuating Responsiveness to Environment factor showed significant loadings on HIT FD (negatively), C, and Sh, and IMPS variables Paranoid Projection and Perceptual Distortion. It is to be noted that all of the variables described here had loadings of .40 or above with their respective factors. The authors concluded that these observed overlaps in common variance are meaningful and, therefore, support the construct validity of the HIT.

In an attempt to develop measures of somatic concern to be used in psychiatric evaluation, Endicott and Jortner (1967) related ratings of somatic preoccupation obtained during interviews to selected HIT indexes of this variable. The subjects in this study were both inpatients and outpatients manifesting a variety of disorders. For the inpatients, the correlations between the number of At responses and the HIT At score and rated somatic preoccupation were .26 and .27, respectively. Animate mutilation and blood responses on the HIT both correlated .21 with rated somatic concern. For the outpatient group, no HIT score correlated with somatic concern. A single measure of somatic concern, derived through a multiple regression equation, combining HIT At, animate mutilation, blood and death responses, and the MMPI Hs scores, was found to correlate .44 and .49 with rated somatic concern in inpatients and outpatients, respectively.

Conners (1965) found significant differences in individual HIT variables and in factor scores when the HIT was administered to disturbed children receiving outpatient care and to a control group of nonclinic children. The disturbed children (neuroses and conduct disorders) received higher scores on R, At, and lower scores on all other variables except V, Sx, Ab, Hs, Pn, and B. Using HIT factor scores, this researcher also found that the neurotic group showed more perceptual differentiation (FA and L primarily) and more inhibition (RT, R, and A negatively) than the conduct problem children. This picture seems quite consistent with usual descriptions of conduct versus personality problem children.

As part of a clinical normative effort, Morgan (1968) provided data on four clinic samples on all 22 HIT variables and compared the results to the appropriate validation sample in Holtzman et al. (1961). The only significant difference (p < .05) found between an elementary school sample and the appropriate comparison group was on V (clinical sample mean = 12.67; standardization sample mean = 5.39). The clinic sample of seventh–ninth graders differed significantly from the norms on A (clinical mean = 25.46; standardization sample mean = 19.23) and M (clinical mean = 36.76; standardization sample mean = 24.37) and these same differences held for a clinic group of tenth–twelfth graders. The only variable showing a significant difference from the norms in the college-clinic sample was I (clinical group mean = 7.22; standardization group mean = 11.08). This study does not seem to provide strong support for the clinical usefulness of the HIT, but several shortcomings of the study should be noted. First, the degree of pathology in the "clinic" sample is questionable since none of the subjects had ever been institutionalized for psychiatric reasons. Second, one can always question the suitability of a standardization sample as a reference group for comparison purposes. There is no guarantee that Holtzman's groups were comparable to Morgan's (1968) in every important respect, even though they were normal and from the same general socioeconomic level.

Krippner (1967) related the reading improvement of 24 second–sixth graders on all 22 HIT variables. Reading improvement was assessed by differences between scores obtained on alternate forms of the California Reading Test administered at the beginning and end of a 5-week reading clinic experience. Four HIT variables correlated significantly (.01 level) with reading improvement: L (.57), Sh (-.60), V (-.96), and Hs (-.65). Since age was not controlled in this study, these correlations are probably spuriously high.

Discussion and Conclusions

The past 10 years have yielded a substantial body of literature relative to the validity of the HIT. As the research stands, it is probably clear that no definite answer could be given to the question "Is the HIT better than
the Rorschach?" With reference to the formal aspects of the HIT—the basic structure of the instrument, the elaborate reference group norms provided, parallel forms—we can agree with Forer's (1965) review in finding the HIT a very attractive research instrument, and, with regard to the latter, many of the personality and developmental studies reviewed here would bear us out. Still, to properly judge the general advantages of the HIT over the Rorschach, many more direct comparative studies are needed.

One can agree with Eysenck's (1965) view that a disproportionate emphasis on reliability as compared to validity is found in the initial construction of the HIT. A more serious attack, however, could be leveled at the lack of concern, or rather, the superficial concern for the conceptual validity of the instrument shown in the introductory HIT monograph (Holtzman et al., 1961). This is not to agree with Eysenck's (1965) condemnation that validity lags so much behind reliability that it "demonstrates pretty conclusively that the underlying notion of the Rorschach test is at fault [p. 217]." This statement, besides being illogical, ignores the issues involved. First of all, it can be seriously doubted that anyone knows for certain what the "underlying notion" of the Rorschach or the HIT is or ought to be. Second, it is questionable that 10 years of research with any instrument, no matter how similar to the Rorschach, can have anything to say about the validity of the theory behind the technique. This is especially true when, as has been the case, many researchers ignore such basic questions as, How should this variable relate to normality-pathology, to dynamics, to overt behavior, to fantasy?

Deficiencies in the designs of specific studies and lacunae in the general areas of research covered by this review have been discussed elsewhere in this article. It will suffice to point out here that with the possible exceptions of developmental studies and those dealing with technological extensions of the HIT, many more studies will be needed in all areas before empirical generalizations concerning this instrument's validity can be offered with any confidence. This is especially true when questions are raised concerning the diagnostic efficiency of the HIT in comparison to that shown by the Rorschach. Besides the dearth of studies of differential diagnosis using the HIT, there are even fewer studies reported in which direct comparisons with the Rorschach are effected.

The scarcity of relevant studies should not detract from the diagnostic potential shown by the HIT as most of the well-designed studies reviewed have produced positive findings. Moseley's (1963) discriminant function analysis (no doubt greatly facilitated by the psychometric characteristics of the HIT) achieved a remarkable degree of diagnostic differentiation. His weights should be further cross-validated, and his approach might serve as a model for future research involving a wider range of nosological groupings. To be sure, this should not be the only approach to research on diagnostic differentiation. Previous critiques of research with the Rorschach by Schafer (1948), Ainsworth (1954), and Holt (1968) have raised important questions concerning the relevance of studies that have been content to relate individual test scores to diagnostic comparison groups. Research validation of these individual scores would, of course, be welcome. But if it is true, as Schafer (1948) and Holt (1968) have asserted, that the clinical use of inkblot perception and other tests involves primarily the description of a subject's personality and behavior in psychological terms, and only secondarily his assignment to a diagnostic category, then the peripheral relevance of score validation to the validity of test interpretation becomes clear. Another issue that confronts the score versus diagnostic category approach is based on the observation, frequently voiced by psychologists and psychiatrists, that protocols of "normal" subjects show a surprisingly high degree of pathology. Needless to say, the claim that the test is at fault when this occurs can easily be countered by the assertions that there is something wrong with nosological schemes, with relating test scores alone to these schemes, or both of the above.

It can be hoped that future research with the HIT will address itself in some way to the issues of personality description versus diagnostic classification and test interpretation versus test score. One solution to these issues might be to pay more attention to nonnoso-
logical criteria (e.g., behavioral rating scales, conditionability scores). To some extent this has been done, and it is interesting to note that in the studies reviewed here, some of the strongest and most consistent findings were observed in studies where HIT variables were related to such nonnosological dimensions as cognitive development, aggression, the process-reactive distinction, and body image.

REFERENCES


HOLTZMAN INKBLOT TECHNIQUE 193


HOLTZMAN INKBLOT TECHNIQUE 193


MEGARGEE, E. I. Undercontrolled and overcontrolled personality types in extreme antisocial aggression. Psychological Monographs, 1966, 80(3, Whole No. 611). (c)


(Received June 30, 1970)