THE EFFECTIVENESS OF CLINICIANS’ JUDGMENTS:
THE DIAGNOSIS OF ORGANIC BRAIN DAMAGE FROM
THE BENDER—GESTALT TEST.1

LEWIS R. GOLDBERG2
University of Michigan

One of the most penetrating criticisms of previous attempts at assessing clinical judgments has been that the experimental predictions asked of the clinician differed in class or in kind from those made in his day-to-day practice. More specifically, the clinician often felt that the prediction of success at flight school (Holtzman & Sells, 1954), competency in clinical psychology (Kelly & Fiske, 1951), or even outcome of psychotherapy (Barron, 1953), called for a range of inferences beyond those usually demanded of him. Some clinicians deny making predictions and, instead, characterize their usual work in terms of “personality description,” “diagnosis,” “dynamic formulations,” or “understanding the patient.” Seen in this light, a good many of the published studies of the clinician’s effectiveness can be dismissed as irrelevant by practicing clinicians who see themselves primarily as diagnosticians or therapists.

To eliminate these objections and thereby place clinicians in their best light, clinical assessment must: (a) involve problems typically encountered by the practicing clinician (for example, diagnostic judgments of neuropsychiatric referrals), (b) allow the clinician to use his favorite techniques (be they tests, interviews, case history material, reports from other services, etc.) in his favorite manner of utilizing them, and then (c) independently validate his conclusions against evidence acceptible to science as a whole; moreover, this entire procedure should be compared with similar judgments made by nonpsychologists (for example, clerical help).

Once one has established the kinds of judgments which clinicians tend to make more validly than less-trained personnel, then the assessment process can be analyzed segmentally to see where and how this increased accuracy comes about. This would involve the same considerations mentioned above, with one important difference: now the clinician would be restricted to one specific instrument or one assessment technique. At this stage one finds many studies aimed at validating a specific technique, yet very rarely do they attempt to approximate the criteria already considered. For this reason, the question of what each diagnostic technique contributes to over-all diagnostic competence lies still unanswered.3

The present paper does not concern itself with the value implications of what the clinician does. Only in the sense that his work is compared with less-trained personnel is any judgment made of the worth of his labors. Whether diagnostic work-ups utilize the clinician’s time to best advantage, whether he should be able to predict overt behavior, whether he should concentrate on research

1 The author wishes to express his gratitude to the personnel of the Ann Arbor VA Hospital for their help in making this study possible. Special thanks are due Philip A. Smith of the hospital and E. Lowell Kelly and Max Hutt of the University of Michigan for their encouragement and criticisms of this paper.

2 Now at Stanford University.
and/or therapy—all such questions are omitted here. The assumption has been made that clinicians spend a good deal of their time in diagnostic testing, and the question asked here is, “how good a job are they doing?”

The experimental answer to this last question hinges on two ancillary ones: who is to be called a clinician, and what is the nature of the independent evidence to test his judgments. The frequent criticism heard when groups of clinicians “fail” in some experimental study is that they were not “expert” enough or that they were essentially “academicians.” Moreover, the criteria typically utilized in studies of this kind are all too often dubious in nature. Even the broad classifications “psychotic,” “neurotic,” “character disorder,” and “normal”—much less the more homogeneous nosological categories such as “paranoid schizophrenia,” “obsessive-compulsive neurosis,” etc.—have no commonly accepted operational definitions. For this reason, it is typical to accept as a criterion either the consensus of many (the majority of whom could be in error) or the judgment of an established few (typically psychiatrists, who are usually responsible for the diagnostic referral to the clinical psychologist in the first place).

In this paper, which reports the first of a proposed series of assessments of clinical practice, the major concern is with diagnosis rather than prediction to later behavior. The nosological category chosen for this study was “organic brain damage, cortical,” because of its inherent criterion for ultimate operational definition; namely, the independent diagnosis of a competent neurological team. The clinicians assessed included those who were currently practicing as diagnosticians (either staff or trainees) at a large VA hospital, and their diagnostic performance was compared with that of nonpsychologists (hospital secretaries). This study was aimed at the second stage of a total assessment project, an appraisal of the validity of a specific diagnostic instrument in the hands of practicing clinicians. Since the Bender Visual-Motor Gestalt Test is the most widely used test for organic brain damage at the installation under consideration (and perhaps at many other installations), this instrument was chosen as the technique in question.

Procedure

Protocols of Bender-Gestalt tests were randomly selected from the files of a VA general medical and surgical hospital. Of these protocols, those from the first 15 patients who had been diagnosed by independent neurological examination as showing clear-cut evidence of cortical impairment were selected to represent patients manifesting organic brain damage (hereafter termed organics). As the nonorganic control group (hereafter termed nonorganics) the protocols of the first 15 patients from psychiatric wards were selected where (a) psychiatric diagnoses were clearly agreed upon, (b) no symptoms usually associated with organic brain damage were reported, (c) no record of cerebral trauma was noted, and (d) any routine examination by the neurological staff was negative to cortical impairment. These latter psychiatric patients were all fairly recent admissions to the hospital at the time they were tested and in general could be characterized as displaying acute rather than chronic symptomatology. Table 1 summarizes certain descriptive variables for these 30 patients.

The 30 Bender protocols were divided into three groups of 10 each, such that in Group I there were 2 organics and 8 nonorganics, in Group II there were 5 of each, and in Group III there were 8 organics and 2 nonorganics. This was done so as to be able to investigate the relationship between frequency of occurrence of a diagnostic entity (base-rate) and accuracy of its diagnosis; also, this allowed the work for each clinician to be broken down into more convenient sections. All protocols were assigned a code number which was
Diagnosis of Brain Damage from Bender

Table 1
Description of Patient Populations

<table>
<thead>
<tr>
<th>Etiology (where available)</th>
<th>Localization (where available)</th>
<th>Symptomatology (where available)</th>
<th>Final Psychiatric Diagnoses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral</td>
<td>Rt. cerebral hemisphere</td>
<td>Convulsions</td>
<td>Paranoid schizophrenia</td>
</tr>
<tr>
<td>Tumor</td>
<td>Lt. temporal lobe</td>
<td>Death</td>
<td>Catatonic schizophrenia</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>Rt. parietal lobe</td>
<td>Chronic brain syndrome</td>
<td>Manic-depressive psychosis</td>
</tr>
<tr>
<td>Thrombosis</td>
<td>Rt. front parietal lobe</td>
<td>Hemiparesis</td>
<td>Character disorder</td>
</tr>
<tr>
<td>Alzheimer's disease</td>
<td>Rt. carotid artery</td>
<td>Hemiplegia, spastic</td>
<td>Conversion reaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Headaches</td>
<td>Anxiety neurosis</td>
</tr>
</tbody>
</table>

Organics
(N = 15; Mean age = 38; Age range = 24-61)

Nonorganics
(N = 15; Mean age = 32; Age range = 23-54)

printed on cardboard and stapled over the patient's name, thus removing this identifying data. On the other hand, all descriptions of how the patient actually drew the Bender designs (i.e., arrows, circles, short notes, etc.) were left on the protocols.

Only the actual reproductions of the Bender designs were used in this study, since uniform elaborations and free associations were not available for all the patients.

Table 2 gives a breakdown of some variables concerning the judges who participated in this project. All of the psychologists were actively engaged in diagnostic evaluations at the time the study was conducted, although they varied both in their general diagnostic testing experience, as well as in their particular experience with the Bender. Many used the test almost routinely as part of their psychological examinations of referred patients, while a few used it only rarely. Without exception, the nonprofessional judges had had no contact with the technique.

Each judge was given the three packets of 10 protocols, one packet at a time in random order. Directions were essentially as follows for all participants:

You will be given 30 Bender protocols for your diagnostic impressions. For your convenience they have been divided into 3 groups of 10 each. Please judge each Bender individually, using any system you normally apply to such a task. Take all the time you want, and feel free to utilize any instruments (i.e., compass, protractor, ruler, etc.) which you feel will increase your diagnostic accuracy.

Please do the very best job you possibly can.5 Record your judgments on the face-sheet attached to each packet.

This face-sheet was a mimeographed form, listing the patients' code numbers in a column down the left-hand side of the sheet. Along the top of the sheet were column headings Organic and Nonorganic, as well as a confi-

Table 2
Description of Participating Judges

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Age</th>
<th>Mean Level of Training in Psychology</th>
<th>Approximate Experience with the Bender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychology staff</td>
<td>4</td>
<td>35</td>
<td>Ph.D. plus 4-10 yrs. experience</td>
<td>6 yrs. 4-9 yrs.</td>
</tr>
<tr>
<td>Psychology trainees</td>
<td>10</td>
<td>27</td>
<td>M.A. plus 1-4 yrs. experience</td>
<td>3 yrs. 1-4 yrs.</td>
</tr>
<tr>
<td>Nonpsychologists</td>
<td>8</td>
<td>25</td>
<td>None</td>
<td>0 None</td>
</tr>
</tbody>
</table>

5 To increase the judges' involvement in this diagnostic task, a bottle of Scotch was offered to the judge who performed most accurately. It was generally felt that all of the judges 'tried their best.' Although the judges spent only 15 to 30 minutes diagnosing all 30 patients, in general they expressed their impressions that they had been as careful in these diagnoses as they would be typically in their regular professional evaluations.
Table 3
Diagnostic Accuracy by Groups of Judges

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Correct</th>
<th>Range</th>
<th>Mean %</th>
<th>Number of Judges Differing Significantly from Chance (50%) at the .05 Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychology staff</td>
<td>4</td>
<td>65%</td>
<td>60-70%</td>
<td>1</td>
<td>25%</td>
</tr>
<tr>
<td>Psychology trainees</td>
<td>10</td>
<td>70%</td>
<td>60-77%</td>
<td>6</td>
<td>60%</td>
</tr>
<tr>
<td>Nonpsychologists</td>
<td>8</td>
<td>67%</td>
<td>57-73%</td>
<td>5</td>
<td>62.5%</td>
</tr>
<tr>
<td>All groups</td>
<td>22</td>
<td>68%</td>
<td>57-77%</td>
<td>12</td>
<td>54.5%</td>
</tr>
</tbody>
</table>

*All mean differences between groups are nonsignificant.

Results

The most striking finding of this study is the complete overlap between groups of judges on diagnostic accuracy (see Table 3). Staff psychologists, psychology trainees, and nonprofessional persons did not differ from each other in their ability to differentiate organic from nonorganic patients by means of their Bender protocols. The judges' degree of successful diagnoses ranged from 57% to 77%, with six judges differing at the .01 level and six more at the .05 level from statistical chance (50%). The remaining ten judges did not differ in their diagnostic accuracy from that attributable to chance alone.

While the Pascal-Suttell Objective Index (1951) was developed to help differentiate psychotic from neurotic populations, it has been used for the diagnosis of organics (Bolland & Deabler, 1956). Cutting scores for this use of the index vary, and extensive normative data is as yet unavailable. If a Z score of 100 is used to separate the groups, the index accurately diagnoses 63% of the patients in this study (Fisher's Exact: \( p = .04 \)). When the cutting score is lowered to 90, the percentage of successful diagnoses increases to 67% (\( p = .01 \)). The optimum cutting score for this population seems to be around 80, at which point the index diagnoses with 80% accuracy (\( p < .005 \)). As the cutting point is lowered below this, accuracy slowly falls away (77%, 73%, and 70% for Z cutting points of 70, 60, and 50, respectively); at a cutting score of 40 or below, only chance results occur.

To check on group differences in interjudge agreement, the percentage of agreement for each group of judges was computed for each patient. The average percentage of agreement over all 30 patients, for each group of judges, ranged from 78% to 85% (chance would predict 50%), with no statistically significant intergroup differences.

Although the groups appeared quite similar with respect to diagnostic accuracy and interjudge agreement, there were large differences between the groups in the amount of confidence they placed in their judgments (Kruskal-Wallis H test: \( p < .01 \)). The non-
professional judges—with no training or experience in Bender interpretation (and therefore no apparent reason for developing confidence in the technique)—were, as a group, much more confident in their judgments than were either the staff or the trainees. And, the trainees displayed more confidence in their diagnoses than the staff. The results present the surprising paradox of an inverse relationship between the amount of experience with the Bender and the degree of confidence placed in diagnoses made from it. Moreover, there was no relationship between individual diagnostic accuracy and degree of confidence.

To test whether the reduced confidence exhibited by the more sophisticated judges was the result of their finer discrimination between the easier and the more difficult judgments (with the more difficult ones being rated with less confidence), the average degree of confidence for each judge was computed for those cases he diagnosed correctly and again for those cases when he misdiagnosed the patient. The difference between these two averages (hereafter termed the index of discrimination), when computed for each group of judges, was found to show no significant intergroup differences. In general, judges were about as confident on cases they misdiagnosed as they were on those they diagnosed correctly. When the index of discrimination was used as the basis for ranking the judges, there was found to be no relationship between this measure and either total degree of confidence or total diagnostic accuracy.

A further measure was developed as an index of clinical judgment,7 under the assumption that a misdiagnosis which had been given with little confidence should not count as heavily against the clinician as one which was given with great confidence. An index was constructed in the following manner: for each correct diagnosis qualified by more than his median degree of confidence, each judge was given two points; for each correct diagnosis given with less than his median degree of confidence, one point; for each misdiagnosis with greater than median confidence, minus two points; for each misdiagnosis qualified by less than median confidence, minus one point. When total scores were computed for each judge, analysis showed no differences between the psychologists, trainees, and the nonprofessional judges on this measure of clinical judgment.

An analysis was made of the judges' tendencies to over-call or under-call organicity in the population they were diagnosing. The trainees as a group made the most "organic" diagnoses, 13.5 per judge, and they were the nearest to judging the actual number of organics in the sample. The nonprofessionals made the least such diagnoses, only 8.5 per judge. These differences almost approach statistical significance (Kruskal-Wallis $H$: $p < .10$), but there was no relationship between the number of patients called organic and diagnostic accuracy. A measure of the dis-

---

7 Suggested by Philip A. Smith of the Ann Arbor VA Hospital.

Table 4

Performance of the Three Groups of Judges on Some Selected Indices

<table>
<thead>
<tr>
<th>Group</th>
<th>Interjudge Agreement Average %</th>
<th>Degree of Confidence</th>
<th>Index of Discrimination</th>
<th>Number called Organic</th>
<th>Index of Clinical Judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychology staff</td>
<td>78%</td>
<td>2.1</td>
<td>1.3</td>
<td>12.1</td>
<td>16.2</td>
</tr>
<tr>
<td>Psychology trainees</td>
<td>85%</td>
<td>2.5</td>
<td>1.1</td>
<td>13.5</td>
<td>20.6</td>
</tr>
<tr>
<td>Nonpsychologists</td>
<td>85%</td>
<td>3.1</td>
<td>1.1</td>
<td>8.5</td>
<td>18.8</td>
</tr>
<tr>
<td>Totals</td>
<td>84%</td>
<td>2.6</td>
<td>1.1</td>
<td>11.1</td>
<td>19.1</td>
</tr>
</tbody>
</table>

---

*a Intergroup differences in degree of confidence significant at the .01 level (Kruskal-Wallis $H$ test); all other differences non-significant.

*b Scored 0 for "blind guess"; 1 for "maybe"; 2 for "think so"; 3 for "fairly certain"; and 4 for "positive."

*c Scored by subtracting the confidence score on misdiagnoses from that on correct diagnoses.

*d The actual number of organics was 15.

*e For method of scoring see text.
crepancy between the number of organics called and the actual number in the sample was computed for each group; neither this measure nor the gross number of organics called was related to any of the previously reported indices. Table 4 summarizes the findings on these various measures.

Since the original material presented to the judges consisted of three packets, each containing different percentages of organic patients, the data could be reanalyzed to investigate base-rate differences on the variables already considered. In this respect, each packet can be thought of as a separate study, each comparable to one conducted at a different type of installation (for example, a GMS hospital, an NP hospital, a home for mentally deficient, etc.). All of the previously reported indices (accuracy, interjudge agreement, degree of confidence, discrimination, clinical judgment, and number of patients called organic) were computed for each group of judges for each of the three different organic base-rates. Table 5 summarizes these results.

The Friedman two-way analysis of variance (nonparametric) was run to test differences attributable to the differing base-rates. The only statistically significant differences uncovered were on the "diagnostic accuracy" and "number called organic" indices. The nonprofessionals were significantly more accurate in their diagnoses as the number of organics in the sample decreased (probably because of their consistent tendency to be the most sparing in their use of the diagnosis "organic"). All groups called more patients organic as the actual base-rates of organics in the sample increased (although in the case of the staff judges, the small size of their group prevented statistical significance).

The Kruskal-Wallis one-way analysis of variance (nonparametric) was carried out to test if any of the differences between mean

---

Table 5

<table>
<thead>
<tr>
<th>Group</th>
<th>Diagnostic Accuracy</th>
<th>Interjudge Agreement</th>
<th>Confidence</th>
<th>Discrimination</th>
<th>Number called Organic</th>
<th>Index of Clinical Judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group I (2 organic; 8 nonorganic)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychology staff</td>
<td>72%</td>
<td>78%</td>
<td>2.2</td>
<td>.2</td>
<td>3.0</td>
<td>7.0</td>
</tr>
<tr>
<td>Psychology trainees</td>
<td>62%</td>
<td>84%</td>
<td>2.3</td>
<td>.6</td>
<td>3.7*</td>
<td>4.9</td>
</tr>
<tr>
<td>Nonpsychologists</td>
<td>78%</td>
<td>88%</td>
<td>3.1</td>
<td>.1</td>
<td>2.0*</td>
<td>9.5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>70%</td>
<td>84%</td>
<td>2.6*</td>
<td>.3</td>
<td>2.6</td>
<td>7.0</td>
</tr>
<tr>
<td><strong>Group II (5 of each)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychology staff</td>
<td>57%</td>
<td>78%</td>
<td>2.1</td>
<td>.5</td>
<td>3.8</td>
<td>3.2</td>
</tr>
<tr>
<td>Psychology trainees</td>
<td>73%</td>
<td>81%</td>
<td>2.5</td>
<td>.2</td>
<td>4.3*</td>
<td>7.2</td>
</tr>
<tr>
<td>Nonpsychologists</td>
<td>68%</td>
<td>80%</td>
<td>3.1</td>
<td>.2</td>
<td>3.0*</td>
<td>5.9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>68%</td>
<td>80%</td>
<td>2.6*</td>
<td>.1</td>
<td>3.7</td>
<td>6.0</td>
</tr>
<tr>
<td><strong>Group III (8 organic; 2 nonorganic)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychology staff</td>
<td>65%</td>
<td>80%</td>
<td>2.0</td>
<td>.7</td>
<td>5.5</td>
<td>6.0</td>
</tr>
<tr>
<td>Psychology trainees</td>
<td>75%</td>
<td>89%</td>
<td>2.6</td>
<td>.1</td>
<td>5.5*</td>
<td>8.5</td>
</tr>
<tr>
<td>Nonpsychologists</td>
<td>56%</td>
<td>89%</td>
<td>3.1</td>
<td>.2</td>
<td>3.5*</td>
<td>3.4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>66%</td>
<td>87%</td>
<td>2.7*</td>
<td>.2</td>
<td>4.8</td>
<td>6.2</td>
</tr>
</tbody>
</table>

* For explanation of indices, see Table 4 or text.
* Differences among nonpsychologists at the three base-rates significant at the .05 level (Friedman two-way analysis of variance).
* Differences significant at the .01 level (Friedman).
* Differences significant at the .05 level (Friedman).
* Differences between groups significant at the .05 level for each base rate (Kruskal-Wallis one-way analysis of variance).

Note.—All other differences nonsignificant.
group scores, at each base-rate, were statistically significant. Just as in the case of the over-all analysis, the three groups of judges were found to differ significantly only in their degree of confidence.

**Interpatient Analysis**

Sometimes the “majority opinion” of a group is more accurate than any of the opinions of its individual members. To check on this, a “group” diagnosis, generated by combining the diagnoses made by all of the 22 judges, was examined for each patient. The amount of agreement between judges ranged from complete agreement on one patient (a manic-depressive psychotic, correctly diagnosed by everyone as nonorganic) to a complete 11–11 split in opinion on two patients (both organics). There was less than 75% agreement on 8 of the 30 patients.

For those 28 patients on whom there were “majority” diagnoses, the group as a whole misdiagnosed two nonorganics (both paranoid schizophrenics) and five organics (of whom three had symptoms which included grand mal convulsions, one was a case of posttraumatic encephalopathy, and one had a deadly glioblastoma multiforme). Thus, if one counts the evenly divided cases as errors (since no group diagnosis was generated), the group as a whole correctly diagnosed 70% of the patients, a figure not significantly different from the mean of the individual diagnoses (68%).

Interestingly, the degree of agreement among the judges did not correlate significantly with the accuracy of their combined diagnoses. Moreover, majority accuracy did not turn out to be related to such variables as the chronological age of the patient nor to his nosological category. On the other hand, since most of the judges tended to underestimate the actual number of organics in the sample, they tended to be most in agreement on patients whom they diagnosed nonorganic (Fisher’s Exact: \( p = .013 \)).

Certain patients must have seemed easier to diagnose than others, since there was a strong relationship between the amount of agreement on a patient’s diagnosis and the total pooled confidence ratings given for this diagnosis (Fisher’s Exact: \( p < .001 \)). Surprisingly, however, the judges were just as confident in rating their incorrect as their correct diagnoses; nor were they any less confident in diagnosing organic than nonorganic patients.

**A Subsequent Exploration**

In the course of examining the relationship between diagnostic accuracy and experience with the Bender, it was noted that the judge who performed most accurately was a trainee who had spent considerable time (as part of his research for a doctoral dissertation) in administering, scoring, and interpreting tests for organic brain damage with a large group of brain-damaged patients. Although a staff judge also had considerable Bender experience and only performed at the median level on this task, one might still wonder whether specific intensive experience with the instrument might not increase diagnostic accuracy. In effect, this hypothesis would imply that although the practicing diagnosticians were not more accurate than nontrained persons on this task, real “experts” with the Bender could surpass them all.

To test this hypothesis, one of the country’s foremost authorities on the Bender test was solicited to take part in this study.* This judge, taking some 20 hours to complete the diagnostic process, did perform more accurately than anyone in the original study—diagnosing 83% of the patients correctly. His scores fell in the middle of the over-all distribution on degree of confidence and discrimination, but he was one of the most accurate in judging the actual number of organics in the sample. He also was at the top on the index of clinical judgment.

Since his performance lends support to the “expert” theory on Bender diagnosis, it seemed legitimate to combine the scores of the top two diagnostic judges in this study into a subgroup of Bender “specialists” and then to reanalyze the data comparing their performance on all the indices with that of the three other groups. This analysis revealed no significant differences on measures other

---

*Special acknowledgment is due Max Hutt of the University of Michigan for the time and thought he invested in this phase of the study.
than those of accuracy and clinical judgment (Fisher's Exact: $p < .004$), on which variables the "specialists" were, in effect, selected.

Discussion

In general, the results indicate that diagnostic accuracy (when using the Bender to diagnose organic brain damage) does not depend on experience or training in psychology (unless, perhaps, that training includes years of intensive work with the instrument in question). If he is not a real expert in the use of the Bender, a clinician will find that his secretary can probably do this particular job of differential diagnosis as well as himself. Moreover, she will most likely have considerably more confidence in her judgments than he would have in his. This makes it all the more unfortunate that one's degree of confidence bears no relationship to his diagnostic accuracy on this task!

Now these results, in themselves, may be embarrassing. And, when one considers the base-rates of organics usually encountered in clinical practice, one might even become alarmed. For in most settings the actual base-rate is closer to Group I (20% organics) than to any of the other groups in this study. And it is in this Group I where the nonprofessionals—by virtue of their tendency to label most patients nonorganic—appear most likely to overshadow their professional employers in diagnostic accuracy (see Table 5). However, neither the nonprofessionals nor the group as a whole did as well as could have been done by merely calling all patients in this group "nonorganic" and thereby diagnosing with 80% accuracy.

Before judging the Bender too harshly, however, the following factors should be considered: (a) the group as a whole did perform significantly above chance (50%) on this task, thus supporting the premise that groups of organics do respond differently to the Bender test than do groups of nonorganics (but when discriminable differences do appear they are typically so obvious that almost everyone can detect them); (b) this study provided no basis for comparing the Bender with other tests for organic brain damage in order to see whether the wide-spread faith in this technique is indeed comparatively justified; (c) it is of course possible that other tests taken in combination with the Bender may permit judges to diagnose more accurately; and (d) one cannot immediately discount the utility of the cues furnished by a face-to-face encounter with the patient, cues which were totally absent in this study.

On the other hand, these results might have been anticipated on the basis of previous work in this area (Bowland & Deabler, 1956). For example, Pascal and Suttell (1951), while formulating their Objective Index to differentiate psychotics from neurotics, have this to say about the differential diagnosis of organics:

The Bender-Gestalt test cannot, in the absence of other data, answer that question (is there cortical damage?) except occasionally in extreme cases which are also clinically apparent (p. 40).

Performance on the Bender-Gestalt test can indicate damage to the cortex only when the damage shows its effect by pronounced disturbance of the ability to execute the test. We know that nine year old children can reproduce the designs without marked deviation from the stimuli. When, therefore, an individual is functioning at a maturational level of nine years with respect to his ability to reproduce the designs, so to speak, we cannot distinguish between his deviation and those of individuals suffering from psychogenic disorders. This fact suggests that damage to the cortex has to be rather severe in its effect on the functioning efficiency of an adult of normal I.Q. before it can be detected by means of performance on the Bender-Gestalt test. This fact also suggests that actual lesions may exist which cannot, on the basis of the deviations noted by us, be detected in performance on this test (pp. 62–66).

Nevertheless, since the evidence suggests the possibility that real experts in the technique may perform with increased diagnostic accuracy on this task, it is conceivable that they might be able to communicate whatever interpretive refinements they possess. In effect, this has been tried with the Pascal-Suttell Objective Index, which at its optimum cutting point performed about as well as did the best of the individual judges. Significant in this connection, however, is the considerably greater length of time taken by the top expert in making his diagnoses as compared to the amount taken by any of the others, and correspondingly the greater length of time
needed to score records by the Pascal-Suttell method. Assuming that real experts—after considerable time and careful scrutiny—can slightly out-perform hospital secretaries, the value of this potential increment in accuracy must be carefully evaluated.

Summary

Staff psychologists, psychology trainees, and nonprofessional (secretarial) judges made diagnostic judgments from the Bender protocols of 15 organic and 15 nonorganic patients and then indicated their degree of confidence for each of their diagnoses. The three groups of judges did not differ in their ability to diagnose organic brain damage from the Bender, although the nonprofessionals displayed considerably more confidence in their judgments than did either of the other groups. The Pascal-Suttell Objective Index approximately equaled the clinical judgments in diagnostic accuracy, but a renowned Bender expert was able to better the diagnoses of the practicing clinicians. The group as a whole diagnosed above a chance level, but when the base-rate of organic patients typically encountered in clinical practice is considered, the results suggest that chances for misdiagnosis could be increased by utilizing the Bender-Gestalt test.

Received January 14, 1958.

REFERENCES


