AN ANALYSIS OF THE MATRIX (PROGRESSIVE MATRICES) TEST RESULTS ON 700 NEUROTIC (MILITARY) SUBJECTS, AND A COMPARISON WITH THE SHIPLEY VOCABULARY TEST.

By H. HALSTEAD, B.A.,

Psychologist to Mill Hill and Sutton Emergency Hospital Neurological Units.

[Received February 16, 1943.]

SEVEN hundred Progressive Matrices records of male neurotic military patients admitted to Sutton Emergency Hospital between April and November, 1942, were compared with a control group (1). The distribution of the patients' scores shows a negatively skewed curve with a clustering of scores below the control median (Fig. 1).

To show the discrepancy more clearly the patients' scores have been divided into percentile grades (Table I). The grade differences are highly significant (p = less than 0·01).

TABLE I.—*Progressive Matrices. Comparison of Grade-percentages, Neurotic Group* (N = 700), *versus Control Group* (= 5072).

Grade.	Control per cent.	Neurotic per cent.
I	5	4·71
II	20	14·86
III	50	37·00
IV	20	31·14
V	5	12·29
	100	100·00

It is recognized that the lower scores obtained by neurotic subjects are partly due to unwillingness to do the test, an " easier " test situation, and to other extra-neurotic factors, but the figures given above suggest that neurotic subjects as a group are somewhat lower in intelligence than the army or civilian population generally. The results also confirmed observations made during test sessions that subjects who scored a high grade on the Matrix Test tended to be less disturbed by their neuroses on written tests of various kinds. Many of these high-scoring patients became interested in the problems in spite of their condition and took pains to arrive at correct solutions. The lower down the scale of intelligence a subject is placed, the less he seems able to deal with this kind of test, and his inability is supplemented by, or reflected in, an unwilling or careless attitude. For this reason, in clinical practice, it is commonly found necessary to check patients' scores by giving them individual tests; this is especially so at the lowest levels, *i.e.*, Grades IV and V.

So far, no reliable method of assessing the extent to which neurotic and other attitudinal factors interfere with " normal " test functioning has emerged.

On the Matrix Test the unreliability of scores is usually estimated by a comparison of scores in individual records with median scores on each of the five sets, A, B, C, D, E, at different intellectual levels. An individual's score is marked " uncertain " if, in any two sets out of the five, there is a difference of three or more points between the number of problems he gets right and the median scores in these sets for an equivalent total score. This discrepancy of three points plus or minus is, of course, arbitrary, and has to be of this magnitude before one can

postulate unreliability, since at all score levels individual score patterns differ. It has been the practice of many testers to credit a subject with his negative score-discrepancies in the sets and thus estimate a so-called " reliable " score, or in other words, the individual's " true level " of ability on the test. This practice is arbitrary and of doubtful validity, as will be shown later.

Since there is a progressive, increasing order of difficulty in the five sections of the test, it is often assumed, quite wrongly, that all reliable records would reflect this by a downward progression of scores from Set A to Set E. The assumption is unwarranted in view of the variations of item difficulty between different individuals at the same or different levels of ability.

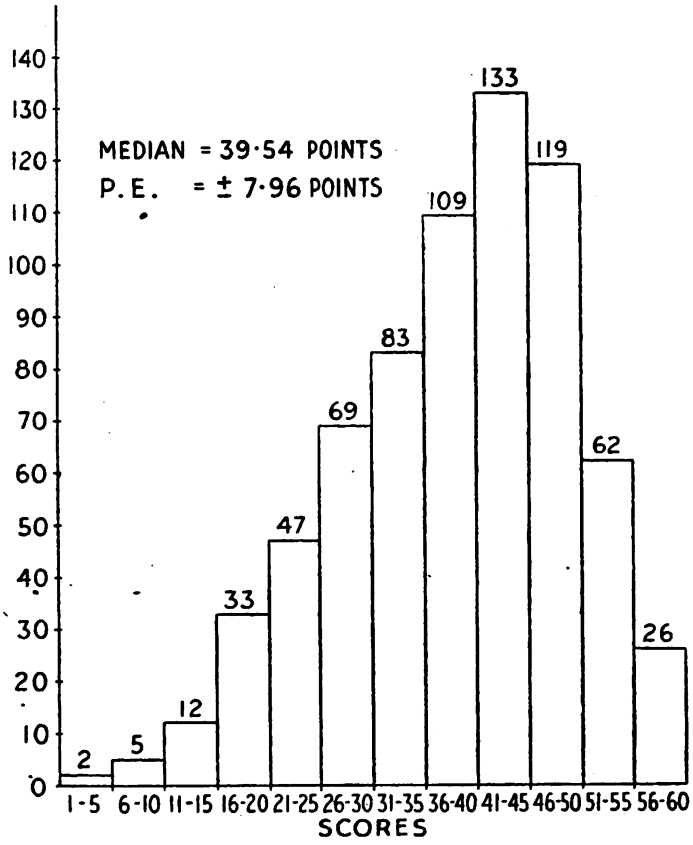

FIG. 1.

In order to test the criterion of unreliability as applied to " uneven " or fluctuating test records, the 700 records of the neurotic patients were divided into two groups, called, for convenience, "straight" and "uneven," the former showing a progressive decrease in scores from Set A to Set E, and the latter showing a reversal of scores between two or more sets. Only 40·4 per cent. (283) fall in the former category, the rest, 49·3 per cent. (417), being " uneven." Of the latter, 84·9 per cent. (354) were " uneven " in one degree, i.e., as between two adjacent sets, and 15·1 per cent. (63) were " uneven " in two degrees, involving three out of the five sets. (This is the maximum number of reversals between sets which can occur.)

In view of this excess of *prima facie* unreliable records, it was decided to make a comparison with an equal number of control records, pairing off score by score throughout the whole range. Thus equal numbers were obtained at each score level in the two groups. The control records, though chosen for comparable scores,

https://doi.org/10.1192/bjp.89.375.202 Published online by Cambridge University Press

were successive, random selections from over 3,000 records. In the control group an even greater proportion of " uneven " records occurred. Only 35·3 per cent. (247) showed a " straight " score progression through the five sets, the remainder, 64·7 per cent. (453), being " uneven." Of the latter, 88·3 per cent. (400) involved two sets and 11·7 per cent. (53) brought in three sets (" doubly uneven "). From these figures it appears that mere fluctuation of scores over the sub-tests is not a criterion of neurosis, for even if we assume that some of the control records came from subjects who subsequently found themselves in a neurological unit, the percentage of fluctuating records far exceeds the percentage of personnel admitted to neurosis centres. Nor is fluctuation a necessary or sufficient criterion of unreliability unless it is substantial, as already mentioned.

It was therefore thought worth while to make a further analysis of the " uneven" records in both the control and neurotic groups, to see if differences in type or degree of fluctuation occurred between them at the various score levels. Examination of the cases which were uneven in one degree only showed that the percentages at each grade-level for the two groups ran closely parallel and were not significantly different (p = 0·95), though there was a significant difference (p = less than 0·01) in the *disposition* of discrepancies over the five sets of the test. At every grade-level, in both control and neurotic groups there were considerably more reversals between Sets C and D than between other sets (Table II), the amount of reversal in a few cases being as much as eight points in both groups.

TABLE II.—*Progressive Matrices. Distribution of Score-reversals.*

(a) *Control Group* (= 400).

Reversals.	Grades. I.	II.	III+.	III—.	IV.	V.	Totals.
A to B . .	5	2	1	4	5	—	17
B to C . .	6	15	45	38	21	10	135
C to D . .	20	55	49	44	52	15	235
D to E . .	3	—	1	—	1	8	13
Totals . .	34	72	96	86	79	33	400

(b) *Neurotic Group* (N = 354).

A to B . .	3	5	11	6	13	2	40
B to C . .	7	23	28	17	10	4	89
C to D . .	18	39	40	47	41	20	205
D to E . .	3	3	2	1	2	9	20
Totals . .	31	70	81	71	66	35	354

The reason for this preponderance will be seen when we consider the order of difficulty of the 60 items in the test. Reversals between Sets B and C are next in order, the highest discrepancy here being five points. There is a tendency for reversals to occur earlier in the test as the intellectual level declines, this being partly a function of the number of items attempted in the different sets. In the neurotic group the percentage of reversals between Sets A and B is significantly higher than in the control group. If therefore unevenness *is* a sign of disturbance on the test, which depends very much on the degree of reversal in score points, then it occurs earlier with neurotics, partly because of their preponderance of lower scores.

In the " doubly uneven " cases the distribution of score reversals shows highly significant differences between control and neurotic subjects (p = less than 0·01). over the score groups. The top two grades approximate each other, but in Grade III the control group has 62 per cent. against 25 per cent. in the neurotic group, with 9 per cent. in Grades IV and V against 43 per cent. neurotics. The control group shows 81 per cent. of discrepancies between Sets B to C and C to D, as against 43 per cent. in the neurotic group, which had 36 per cent. A to B and C to D discrepancies, again indicating an earlier " disturbance." There was no significant

https://doi.org/10.1192/bjp.89.375.202 Published online by Cambridge University Press

difference in the *amount* of reversal between the two groups either for single or double discrepancies, so this factor does not help us *per se* to single out " neurotic " records.

A comparison of the percentages of all the " uneven " records at different score levels as between control and neurotic groups shows significant differences (p = less than 0·01) in the interval 41 to 45 ; the controls show 84 per cent. reversals as against the neurotics' 60 per cent.

The Time Factor.

The records of the control group were obtained under a time limit of 40 minutes but the neurotics had no time limit, though individual times were taken for most of the records. This affects some of our comparative figures to an unknown but probably small degree (*cf.* below). It was thought that better results might accrue from a suspension of the time limit with these subjects, who show a wide range of emotional disturbances. It is interesting to note that when 230 neurotic subjects were given the Matrix test by highly trained sergeant testers with a 20-minute time limit the scores obtained were somewhat higher than those reported above. It may therefore be advisable to time group tests even under clinical conditions, and only if this is done can there be a valid comparison with " normal " records. The imposition of a time limit may help to remove distractability, loss of interest, etc., provided there is a good atmosphere for the test.

In this section therefore only the neurotic records can be dealt with. In the " straight " records the correlation between time and score is little more than zero, with no significant differences between the median scores nor between interquartile ranges for different time values. In the " uneven " records there is a slight tendency for test times to increase as the intelligence level rises, but this is not significant. There is also a tendency for Grade I scorers to show a wider variation in times taken to do the test—again not significant. The median times and interquartile ranges agree closely between " straight " and " uneven " records. On the total group the Median Time is 43 minutes and the Interquartile Range 27 minutes. The shortest times recorded were between 10 and 20 minutes, with a fairly normal scatter over all score levels. The longest time taken by any subject in the " straight " group was 2 hours. Six people in the " uneven " group exceeded this, the longest period being 2 hours 45 minutes.

These results confirm the numerous reports that quickness and ability (or " speed " and " power ") on mental tests are correlated, since the average times vary little over an *increasing* score range. If we calculated a time score index for each subject we should, under normal conditions, have an indication of efficiency, as suggested by Babcock and others (2). When large numbers are being tested with a time limit, it is not practicable to record individual times, though such an index would be a useful adjunct in assessing individual cases clinically. Since we have no control figures we cannot make reliable inferences as to the relationship

Table III.—*Progressive Matrices Time-Score Indices (Average Minutes per Item) (N = 676, Neurotic Group).*

Score group.	Midpoint.	N.	Centiles. 25th.	50th.	75th.
56 to 60	57·5	26	1·23	0·83	0·55
51 ,, 55	52·5	62	1·13	0·91	0·64
46 ,, 50	47·5	113	1·17	0·87	0·66
41 ,, 45	42·5	132	1·40	1·06	0·81
36 ,, 40	37·5	103	1·48	1·14	0·74
31 ,, 35	32·5	80	1·54	1·21	0·95
26 ,, 30	27·5	65	2·06	1·32	0·88
21 ,, 25	22·5	45	2·46	1·75	1·13
16 ,, 20	17·5	32	3·26	2·57	1·53
11 ,, 15	12·5	11	—	—	—
6 ,, 10	7·5	5	—	—	—
1 ,, 5	2·5	2	—	—	—
		676			

https://doi.org/10.1192/bjp.89.375.202 Published online by Cambridge University Press

between neurosis and efficiency (in the present sense) on the test, although we know that neurotics as a group are less efficient than normal individuals. We have, however, information as to speed of functioning in the test, and this, together with qualitative remarks about individual attitudes during the test, can be very helpful clinically. For this reason we append a table (Table III) showing the Medians and Quartile Deviations in time-per-item at different score levels in our neurotic group. The time devoted to the different items of the test varies greatly and, other things being equal, is a function of the difficulty of the items to each individual. Item analysis shows different overall difficulty orders at different score levels. The composite difficulty order is shown in a later section.

When we plot time against age, again in the neurotic group only, we find hardly any correlation either in the " straight " or in the " uneven " records. The Median times and P.E's. fluctuate over the age groups, with a very slight tendency for times to increase with age. There is also a slight but insignificant tendency for a wider variation in test times to occur in the " over 30's."

Age.

The age distribution of the neurotic subjects shows a marked difference from that of the controls. Whilst 88·6 per cent. of the latter were in the 21 to 30 age group, the corresponding figure for the former is 52·5 per cent., with 35 per cent. for ages 31 to 40. There was a lag of a few months between the periods over which the two groups were tested, but this was not long enough to account for the difference, which is significant. The Median Ages for the two groups are 25·82 and 28·30 respectively. When the proportions of " uneven " records for the two groups were plotted against age levels, the differences between the groups follow the total age distributions and are therefore significantly different, but the proportions at different ages are not; the expectation of reversals between sets or sub-tests is about the same at each age from 21 to 45–50.

The records were next analysed in respect of score versus age. In both groups the correlation is little higher than zero and therefore not significant. However, by distributing scores in grades and using χ^2 as a test of significance, certain differences appeared. In the control group, a comparison of the two age groups, 21 to 25 and 26 to 30, showed a significant decrease in score with the older group (Table IV).

Table IV.

Age.	Straight records.		Uneven records.	
	Median score.	P.E.	Median score.	P.E.
21 to 25	45·0	17·6	42·3	12·4
26 ,, 30	32·3	20·3	37·4	14·7

In the neurotic group there appears to be a slight downward tendency in score with increasing age from 20 to 50, but not high enough to be significant. The highest drop is between age groups 31 to 35 and 36 to 40, but as the numbers in the latter group are small ($N = 35$), we cannot assume reliability.

Attitudes to the Test.

On the whole, the Matrix Test is well taken, even in a neurological unit. A great deal depends upon the conditions under which the test is given. If it is given soon after admission to hospital, the results are adversely affected. The patients have not had time to settle down in the new environemnt. They may have travelled long distances and arrived tired, all this in addition to the symptoms for which they are referred. They may also have had preliminary interviews, blood tests, etc. It is advisable, if conditions permit, to postpone testing for a day or so. Also, much depends upon the test atmosphere, the tester, the kind of encouraging talk given before the test, the room, lighting, and so on.

For this section several thousand Matrix records from neurotic patients were examined. The tests were administered by different people, including psychologists, sisters, and nurses. No analysis has been made of scores obtained under different testers. Each person, however, administered the test along broadly similar lines,

with individual modifications, and the tester was instructed to make notes on the record of each subject who showed a negative attitude towards the test. No remarks were made if the attitude was positive or neutral. The number of recorded cases depends upon the tester and his own attitude to the situation, the amount of his interest, observation, and other personal factors. However, every tester supplied comments incorporating his or her own observations and remarks made by patients. Of 2,500 records taken at random, only 5 per cent. (126) carried remarks on negative attitude. These were analysed with reference to complaints, score, age, and time. The complaints fell into the following categories (Table V):

TABLE V.

	Number of patients.
A. *Negativism, or "Won't" Attitudes*	
i. Resentment, aggression, sullenness, etc.	11
ii. Test regarded as a joke, as childish or humiliating; "can't be bothered" attitude, etc.	13
B. "*Can't*" *Attitudes.* The main theme in this group was one of inability to concentrate; this applies to all in the group, though the causes differ. The categories appearing are:	
i. Anxiety, agitation, tremors, restlessness, giddiness, twitching, palpitations	32
ii. Vision: eyestrain, blurring and shifting of test material, spots before the eyes, double vision, distortions .	26
iii. Headache	24
iv. Poor comprehension of task	12
v. Depression, confusion, weeping	5
vi. Collapse	1
vii. Others	2

Most of the subjects managed to struggle through the test, but a few were unable or unwilling to continue and left the test unfinished. Occasionally one would come across amusing remarks: "It is nearly sending me balmy"; "This is ridiculous; it makes me feel humiliated"; "I don't want my brain tested"; "Next time I have anything like this to do, I shall start at the end and work backwards,"; "This is going back to infancy; I get optical illusions"; "The patterns are going round and round." A question commonly asked was, "Why are we having this; do they think we're mental?" As one would expect, the scores obtained from these patients are much lower than those of the patient groups as a whole. The median score is 19·8, as against 39·5, with a Probable Error of 18·3 as against 15·6 points. In view of these attitudes towards the test, it is difficult to estimate the reliability of the results, though with such a substantial and highly significant discrepancy one is inclined to infer that most of the complaints came from the dull and backward subjects. A division into the conventional score categories gives the following distribution (Table VI). The normal or expected percentages are given in brackets for comparison. It will be seen that over 70 per cent. of the records are in Grades IV and V, with less than 9 per cent. in the two top grades. This confirms earlier remarks on the better attitudes of higher grade subjects.

TABLE VI.—"*Complaints*" *Group. Distribution by Grades.*

Grade.	Percentage.	Percentage.
I	0·8	(5)
II	7·9	(20)
III	19·8	(50)
IV	36·6	(20)
V	34·9	(5)
	100·0	(100)

https://doi.org/10.1192/bjp.89.375.202 Published online by Cambridge University Press

Impressions received during test sessions had been that, on the whole, most of the complaints about the test came from older men, but the distribution of the 126 records shows no significant differences from the total neurotic group. The median age is 28·2 and the interquartile range 11·9 points (as against 28·3 and 10·3 respectively for the neurotic group as a whole). Moreover, the test times do not differ significantly for the " complaining " subjects. The Median Time is 46 minutes and the interquartile range 23 minutes (as against 43 minutes and 27 minutes respectively).

Estimating the Reliability of Scores.

An analysis of the Median scores and Interquartile Ranges in each of the five sets of the Matrix Test at different intellectual levels shows variations in " scatter." For this reason the present method of assessing unreliability and of estimating probable " reliable " scores is open to question, and can at any rate be improved.

The amount of trouble one is prepared to take to assess unreliability and to estimate " reliable " scores depends upon one's interest in individual cases. In clinical practice it is often deemed more convenient to use such statistical aids as are available rather than retest. For special purposes the administration of the test individually, with unobtrusive timing of items, getting the subject to verbalize his thought processes, noting his wrong choices, and in some cases taking him back through his errors, is the best procedure. A retest under group conditions may produce a different pattern of scores but a still unreliable result. However, in these circumstances one often feels justified in summing the correct answers from both sessions to get a better result.

A first step in assessing a record would be to see whether the scores in each set, when compared with group figures for the same total score obtained by the subject, fell inside or outside either the middle 50 per cent. (P.E.) or some other interval, e.g. standard deviation. A score which is within these " normal limits " would be deemed reliable, and a score outside would be regarded as unreliable. A second step would be to compare the total record with a list showing the order of difficulty of items through the whole test. Table VII shows such a difficulty order. This was obtained by plotting the frequencies of correct answers for the 60 items from 2,790 unselected control records.

If in a given record a fair number (to cover chance successes, which are in the ratio of 1 to 7 for each item over the test as a whole, and individual differences in relation to item difficulty) of successes on more difficult items appear *pari passu* with a fair number of failures on easier items, one can assume unreliability and make an assessment of the probable reliable score. Each case would have to be taken on its merits and the assessment would depend to some extent upon the tester, but he would be in possession of additional information (the difficulty order of items) to help him form an estimate. It is not possible to establish fixed criteria for this second step because of the varying individual score patterns. Also, the method presupposes that an appreciable number of items have been attempted on two or more sets of the test. It should be of service down into Grade V intellectual level. The reliability of estimates by this method of item comparison increases with the number of items attempted. Obviously, if only one set is completed, or five sets are barely attempted, any system of estimating probable scores on the test breaks down. Each of the five sets of the Matrix Test follows a logical principle, and the author of the test has made the sets as a whole increase in difficulty from Set A to Set E. There is also a fairly uniform progression within each set. Because of this progression and its particular practice or learning sequence, the use of the over-all order of difficulty is somewhat mitigated, but as far as one can judge there seems to be a strong tendency on the part of most testees to approach each item as a separate problem.* If therefore displacements of items on a record, as compared with the above rank order, are of more than one place, the judgment of unreliability is greatly strengthened, since the differences in frequency of correct scores is discriminative even between adjacent ranks. The greater the displacement of any item in either direction, the greater its unreliability as shown by rank frequencies (i.e. the number of subjects scoring correctly on the various items).

* *Cf.* Section on "Analysis of Errors."

TABLE VII.—*Progressive Matrices. Difficulty Order of Items (N = 2,790 Controls).*

Number of item.	Frequency of successes.	Number of item.	Frequency of successes.
A1	2,790	B7	2,117
A2	2,790	D8	2,050
A3	2,788	D7	2,027
A4	2,787	E1	1,978
A5	2,783	D9	1,962
A6	2,751	C8	1,861
B1	2,737	C6	1,832
B2	2,729	C9	1,829
A9	2,694	A11	1,759
D1	2,662	B12	1,666
B3	2,658	B10	1,646
C1	2,643	E3	1,611
A8	2,631	E2	1,599
A7	2,592	B9	1,471
B8	2,567	B11	1,437
B4	2,550	D10	1,417
A10	2,539	C10	1,303
C2	2,538	C11	1,094
B5	2,517	E5	1,061
A12	2,457	D11	977
C3	2,448	E4	931
D5	2,442	E9	886
D2	2,410	E6	850
C7	2,406	E10	581
D3	2,384	E11	496
B6	2,293	D12	476
C5	2,278	E7	281
D6	2,210	C12	242
D4	2,190	E12	157
C4	2,167	E8	129

Positions of Items in Standard Sequence.

The position of each item, as above, is shown in brackets against its position in the standard sequence.)

Set A.	Set B.	Set C.	Set D.	Set E.
1 (1)	1 (7)	1 (12)	1 (10)	1 (34)
2 (2)	2 (8)	2 (18)	2 (23)	2 (43)
3 (3)	3 (11)	3 (21)	3 (25)	3 (42)
4 (4)	4 (16)	4 (30)	4 (29)	4 (51)
5 (5)	5 (19)	5 (27)	5 (22)	5 (49)
6 (6)	6 (26)	6 (37)	6 (28)	6 (53)
7 (14)	7 (31)	7 (24)	7 (33)	7 (57)
8 (13)	8 (15)	8 (36)	8 (32)	8 (60)
9 (9)	9 (44)	9 (38)	9 (35)	9 (52)
10 (17)	10 (41)	10 (47)	10 (46)	10 (54)
11 (39)	11 (45)	11 (48)	11 (50)	11 (55)
12 (20)	12 (40)	12 (58)	12 (56)	12 (59)

It is interesting to compare the standard sequence of the items in the Matrix Test with the over-all difficulty order obtained. If we plot the two sequences against each other by sets of 12 items (Table VIII), we can see the amount of displacement in the five standard sets in relation to total difficulty of items :

https://doi.org/10.1192/bjp.89.375.202 Published online by Cambridge University Press

TABLE VIII.—*Progressive Matrices. Standard Sequence Compared with Difficulty Order Divided into Five Equivalent Sets (Successive).*

Sets.	1.	2.	3.	4.	5.	Total.
A	**7**	3	1	1	0	12
B	4	**3**	3	2	0	12
C	0	2	**3**	6	1	12
D	1	4	4	**1**	2	12
E	0	0	1	2	**9**	12
	12	12	12	12	12	60

It will be noticed that Sets A and E, the easiest and the hardest, show the least amount of displacement of over-all difficulty. Most subjects are capable of doing all items in Set A, so that here the real difficulty order is less certain than with later items. It would be more certain perhaps if frequencies of earlier items were plotted on an isolated sample of low-grade subjects or young children. Set D shows the greatest amount of displacement, with sets B and C lying between. The table also shows clearly why there is a tendency to more reversals between Sets C and D, as already mentioned.

ANALYSIS OF ERRORS (PROGRESSIVE MATRICES).

Control records (N = 2,790) were analysed for frequencies of errors. Table IX shows the most frequently chosen wrong items. The choices depend upon several factors, such as level of ability, attitude, number of choices, possibility of choice by elimination of items already used in the pattern, and many others.

TABLE IX.—*Progressive Matrices. Wrong Items most Frequently Chosen.*

Set A.	Choice Number.	Set B.	Choice number	Set C.	Choice number.	Set D.	Choice number.	Set E.	Choice number.
1	–	1	1	1	2	1	4	1	1
2	–	2	2	2	3	2	3	2	3
3	3	3	2	3	6	3	4	3	2
4	3	4	4	4	4	4	5	4	3
5	2	5	5	5	4	5	2	5	7
6	2	6	2	6	8	6	7	6	8
7	4	7	6	7	4	7	8	7	7
8	6	8	2	8	6	8	1	8	3
9	5	9	1	9	1	9	3	9	6
10	6	10	1	10	8	10	5	10	2
11	1	11	3	11	5	11	6	11	4
12	1	12	2	12	3	12	8	12	–

"Choice number" refers to the number of one of the "pieces" at the foot of each page of the test.

In Set A all the errors are errors of perception, this set being more of a perceptual than a conceptual test, i.e. eduction of relations. Some of the errors are similar in pattern to the correct items and may have been committed through visual errors, carelessness, haste, etc. Similar remarks apply to the errors involving reversals—of whole figure, of lines, of figure-ground. Some are quantitative errors, e.g. too many or too few lines or dots. Set B shows only five "perceptual" errors (it is in this set that a change-over to eduction of correlates takes place). These again include reversals and errors of size and shape. In eight cases the choice (error) is a copy of an adjacent (discrete) part of the pattern. Inadequacy of reasoning begins rather late in this section (about B9). In Set C the number of perceptual errors is reduced to two. There are four instances of copying adjacent fragments of the design, two which might loosely be termed "clang" completions, e.g. the A.B–A sequence, as in simple tunes, and four of incomplete inference, the final step having evaded the subjects. The perceptual errors in Set D are again only two (three) in number, with three copies of neighbouring items, four "clang" completions, and four really inadequate choices with probable guessing. Finally, in Set E,

https://doi.org/10.1192/bjp.89.375.202 Published online by Cambridge University Press

there is only one error which might be called a perceptual error, one " clang " choice, one absolute failure (E12), the rest being failed chiefly through inadequacy of the reasoning process. There is a tendency to complete this process by elaboration, i.e. by choosing an alternative which is more elaborate than any already in the pattern—an assumption that the reasoning necessarily involves a progression of items from simple to complex.

With the lower-grade subjects, who cannot deal simultaneously with more than a very small number of variables, e.g. two or three, the errors are chiefly " perceptual," the next most frequent being due to choosing an item like the one adjacent (to the blank space); once this happens, many succeeding items are automatically failed by this follow-my-leader procedure. When subjects are tested individually it is possible to avoid some of these errors, not so much by prompting as by issuing warnings at intervals.

Most of the errors of the higher-level subjects are due to inadequacy of the reasoning process, with completion often by a " hunch." Mathematically-minded subjects seem to do as well as any on the test, and indeed some items in Set E can only be solved logically. High scores have, however, been obtained by artistic people who have an eye for form (Gestalt), symmetry, etc.

Much can be learned about a subject's attitude and mentation from a scrutiny of his errors, though this is time-consuming and can be justified only in special cases. The present, rather sketchy analysis is put forward as a supplement to information already published about this test.

Comparison with Shipley Vocabulary Test.

Of the Matrices records on neurotic subjects, 650 had comparative records on the Shipley Vocabulary Test, which was used in a slightly revised form, involving one or two minor alterations of the alternative (selective) words. Also the time limit was increased from $7\frac{1}{2}$ to 10 minutes, to allow subjects to write their choice-words on a slip, instead of underlining them or giving the column number. This was necessary owing to paper shortage, and to avoid the confusion which might arise from recording column numbers on wrong lines on the slip. The modification affects the comparison with results on the older form only slightly. The increased time limit is adequate in the majority of cases. Most of those who fail to attempt all the words would be unable to do so on the original method. The depressed patients usually have adequate time to read the whole list of 40 words, and the low scorers can deal with all the words they know.

For comparison with service norms it was possible to bring in 1,326 Shipley records on neurotics. The comparative figures are shown in Table X. It will be noticed that here again there is a skewness towards the lower levels, but less so than with the Matrix Test. A comparison of the neurotic records with those of the control group grade by grade shows a less significant difference than with Matrices ($p = 0{\cdot}25$).

Table X.—*Shipley Vocabulary Test. Comparison of Grade-Percentages. Neurotic Group* ($N = 1326$) *versus Control Group* ($N = 5000$).

Grade.	Control percentage.	Neurotic percentage.
I	10	10·41
II	20	12·30
III+	20	19·70
III—	20	17·20
IV	20	25·67
V	10	14·72
	100	100·00

N.B.—In this table different gradings have been used from those shown in preceding tables.

In neurotic subjects the Shipley Vocabulary Test " holds up " comparatively well. There is little discrepancy in the two middle grades which make up 40 per cent. of the scores; the displacement is from the upper to the lower grades. Every

patient is able to comprehend the task and, except in the lowest 5 per cent. of cases, where the score reached is only about seven words out of a total of 40, there is a fluctuating but linear progression throughout the whole range of scores.

The correlation obtained with Progressive Matrices was $0{\cdot}547 \pm 0{\cdot}021$. This is as high as one would expect in view of the different factor loadings of the two tests.

The regression lines show a linear regression of the Shipley Test on the Matrix Test, with a slightly curvilinear regression of the Matrix Test on the Shipley Test at the lower score levels. The number of cases at these levels is small and therefore unreliable, but the curvilinearity suggests that a certain minimal level of old-established ability, as indicated by the Vocabulary Test, is needed before one can get into the stream on the Matrices Test; in other words, the bottom 5 per cent. or so on the Shipley Test do less than proportionately well on Matrices, on the average.

TABLE XI.—*Comparison of Progressive Matrices and Shipley Vocabulary Scores on Neurotic Group* ($N = 650$).

Shipley Grades	Matrices grades. I.	II.	III.	IV.	V.	Total.
I	**23**	20	24	2	0	69
II	17	**39**	57	9	5	127
III	22	52	**124**	42	18	258
IV	3	14	46	**46**	22	131
V	0	1	20	15	**29**	65
Total	65	126	271	114	74	650

Grade discrepancies. Number of grades discrepant.	N.	Percentage.
"Shipley" higher: 4	0	30·6
3	7	
2	51	
1	141	
Agree	261	40·2
"Shipley" lower: 1	130	29·2
2	56	
3	4	
4	0	

Table XI shows the Shipley Test plotted against the Matrix Test in grades. From this the discrepancies can be seen clearly. The discrepancy curve is almost normal. A conventional grading used for each test embraces 30 per cent. in Grades I and II, 40 per cent. in Grade III, and 30 per cent. in Grades IV and V. The discrepancies between the two tests are similarly distributed, i.e. 30·6 per cent. of total scores favour the Shipley Test, 40·2 per cent. agree as to grade, and 29·2 per cent. favour the Matrices Test; 69·4 per cent. of the discrepancies between the tests are misplaced one grade only. Using the Shipley Test as "standard" and taking each of its grades separately, we find that Grade III (the median score levels) shows the lowest percentage of scatter on Matrices (52 per cent.), and Grade I the highest (69 per cent.). If the Matrices Test is taken as a standard, Grade I scorers show the lowest percentage of variance on the Shipley Test, and Grade II the highest. From a purely positional point of view, one would expect to find the least scatter on Grades I and V, which can vary in one direction only, and most on Grade III, which can trespass into two grades, i.e. up or down. But whether we take the Shipley Test or the Matrices Test as standard, the variation shown on the correlative test at any grade level differs little from equal expectancies for each grade ($p = 0{\cdot}08$). A comparison of the amount of scatter on Matrices shown by each grade of the Shipley Test shows the one significant difference to be in Grade I, with a higher percentage than one would expect to be caused by chance. From this result one is tempted to infer that the top scorers on the Vocabulary Test (the

verbally facile subjects) are more disturbed by a test such as Matrices than the top scorers on Matrices are on the Vocabulary Test.

The correlation between age and score on the Shipley Test is positive but near to zero, with a slight tendency to higher scores and less variation with increasing age from 20 to 45 (50). This tendency, though not statistically significant, suggests a better stability on this test with increasing age, whether it is in attitude or in mental (verbal) organization.

Mental Deterioration.

There seems to be adequate evidence that there is a high correlation between vocabulary scores and scores on tests of " general intelligence," but most of the correlations reported are with verbal intelligence tests which include a high " V " loading. It has also been commonly accepted that in tests of deterioration the vocabulary *score* is to be used as an index of the pre-deterioration level of the patient, the argument being that a person's vocabulary is an old and early established mental habit and that it " holds up " well with age. Both these postulates are true, and the writer's own work (unpublished) on seniles shows that a score on a vocabulary test, even of the " inventive " kind where the subject has to supply his own definitions, is sometimes possible when many other tests are too difficult, even in cases where psychomotor activity is the only kind that can reliably be measured, or down to the lower limits of verbal comprehension. But in view of the wide individual differences of mental functioning, as shown by innumerable " profiles " of abilities obtained on test batteries, the discrepancy between scorers on combinations of tests and scores on a vocabulary test has to be quite marked before one can begin to suspect deterioration (3).*

The present report shows the danger of inferring deterioration from such discrepancies too readily, since almost as many subjects score lower on the Vocabulary (selective) Test compared with the Matrix Test as *vice versa*, some of the discrepancies being considerable. The same kind of thing happens with " normals." Many of the investigations on vocabulary discrepancy scores showing deterioration report group averages, and show wide deviations, indicating the breadth of individual differences. Substantial discrepancies are obtained on subjects already known to be seriously deteriorated or demented, either through natural senility or organic disturbance. In any case, one would require additional knowledge such as school and job records in order to be able to assess the pre-deterioration level. Babcock (2) has concentrated on " border-functioning " subjects in an endeavour to assess deterioration in its infancy, and she finds a relationship between score patterns (on her Deterioration Scale) and certain types of clinical patients. The present writer would like to see, in addition to thorough analyses of as many reputable tests as possible, results on representative samples of the population, between the ages, say, of 20 and 80, both on separate tests and on selected groups of tests. Wechsler (4) with his " Belle-vue " Intelligence Scale, has shown the way in this respect, since he provides figures from the age of 10 to the age of 60, and has given useful information on the normal decline of intelligence and on different ability patterns on the test at different age levels. The Matrix Test is useful in that the age range for testing is from 10 or less right through to old age. When such investigations have been done, clinicians will find themselves in a better position to attempt the assessment of mental deterioration. It is imperative that we should know the limits of normal variations as between different tests, including vocabulary tests, over a wide range of age and intelligence.

Discussion.

At the present time, when mental testing is being done on a scale unprecedented in this country, both in the Services and in the neurosis centres, a wealth of information is being accumulated in the form of test records. There is now a substantial and increasing number of reputable, well-standardized, and home-produced tests in use, especially in the Services, and psychologists have an unprecedented opportunity of analysing these tests and getting the most out of them. In the clinical field, the

* Brody takes into account the quality of the vocabulary responses as well as the score.

personnel using such tests will appreciate as much information as can be obtained to supplement the scores and such observations on test behaviour as they can make on the spot.

The present article is an attempt to add to the existing knowledge of two widely used and well established tests, and it is hoped that the findings will be of direct use to psychiatrists, psychologists, and others concerned in mental testing.

Summary.

1. Seven hundred Progressive Matrices records from neurotic military patients were compared with controls. It was found that the neurotics as a group had significantly lower scores. Subjects at higher intellectual levels seem to take written tests better than the low-grade subjects, who often need to be re-tested individually.

2. When the records were divided into " straight " and " uneven," the latter preponderated. The control group showed an even higher percentage of " uneven " records, so that mere unevenness between the five sets is no criterion of neurosis. There was no significant difference in percentage of uneven records between different score levels in either control or neurotic groups, nor did the two groups differ significantly in amount of reversals. The neurotics showed a tendency to reversals of scores earlier in the test, partly because of their lower scores. In both groups there were significantly more reversals between sets C and D.

3. An analysis of the time factor in the neurotic group, who were tested without time-limits, showed a very low positive correlation between time and score, with no significant differences between " straight " and " uneven " records, but with a slight tendency for time and score to vary directly in the latter, there being more variation in Grade I scores. The Median Time over the whole group was 43 minutes, with a P.E. of 27 minutes. Since times do not alter significantly with *increasing* scores time-per-item decreases as score-level rises, i.e., quickness and ability are positively correlated.

Time and age showed a very low positive correlation, with a slight tendency in the " uneven " group for times to increase with age and for scores over 30 to show more variation.

4. The neurotic group showed a lower median age and wider range than the control group. The correlation between age and score is almost *nil*. The percentage of " uneven " records does not differ significantly from chance (equal) expectations at different ages in either the control or the neurotic groups, though the former shows a sigificant difference between *two* age groups, i.e. 21–25 and 26–30, the latter ages giving lower scores. There is a *very* slight tendency in the neurotic group for scores to decline with age, from 20 to 45.

5. Attitudes of neurotic subjects to the test are on the whole good, only 5 per cent. of a sample (2,500) of records showing really negative attitudes. These are analysed into various categories. There is no significant difference in age or times taken among this " complaining " group, but the low scorers show poorer attitudes towards the test. The need for good testing conditions is emphasized.

6. The current method of assessing unreliability of scores on the Matrix Test is mentioned, with further suggestions incorporating a comparison of individual records, with an over-all order of difficulty of the 60 items on the test and an index of efficiency by the use of a time-score index. A combination of methods gives the best assessments.

7. 2,790 Matrix (control) records were analysed for frequency of wrong choices. It is shown that these follow the progression of the test, i.e. broadly from perceptual to conceptual. There is a tendency for a particular wrong approach to persist amongst low-scorers through several items. Types of errors in each of the five sets are mentioned.

8. A comparison is made between 650 Shipley and Matrix records showing a normal curve of grade-displacements between the tests. The majority of these displacements are of one grade only.

9. The use of a vocabulary *score* in estimating mental deterioration is mentioned, and the danger of making facile inferences from discrepancy scores (vocabulary scores minus scores on other tests) is emphasized. There is a need for information

about the performance of subjects over as wide an age range as possible on the more reputable tests now in use.

I wish to express my thanks to Dr. Louis Minski, Medical Superintendent, Sutton Emergency Hospital, and to Dr. W. S. Maclay, Medical Superintendent, Mill Hill Emergency Hospital, for the use of clinical material, also to the Director of Selection of Personnel, War Office, for the use of data for comparison.

References.

(1) Raven, J. C., " Standardization of Progressive Matrices," *Brit. J. med. Psychol.*, **19**, pt. 1.
(2) Babcock, H. (1941), *Time and the Mind*, Sci.-Art Publishers, Cambridge, Mass.
(3) Brody, M. B. (April, 1942), "The Measurement of Dementia," *J. Ment. Sci.*, **88**, No. 371.
(4) Wechsler, D. (1942), *Measurement of Adult Intelligence*, Williams & Wilkins Co., N.Y. (2nd edition).