TESTS FOR CREATIVE ABILITY IN MACHINE DESIGN

ANNUAL REPORT

Project: ONR: 458: MEY: al Nr 192-034

Contract: Nonr - 530 (02)

Submitted by

C. F. Schumacher

G. R. Maxson

H. Martinek

NOV 15 1953

TABLE OF CONTENTS

TABLES

I. INTRODUCTION

From the industrial point of view, the efficient selection of personnel becomes increasingly more important as the complexity of the job increases. Mistakes made in the selection of punch press operators are much less costly and easier to correct than mistakes made in the selection of vice presidents. A somewhat similar situation exists in the field of education. The selection of the wrong candidates for specialized training is costly, in terms of time and effort wasted, both to the student and to the institution whose facilities and staff are limited.

One area in which the selection of personnel capable of success on the job is extremely critical at both the college level and in the field is the area of machine design. Able designers are few, and, especially in times of national emergency, there is an urgent demand for their skills. In such a situation, even slight improvements in the prediction of success on the job or in training may yield substantial dividends in terms of types, numbers, and quality of products developed.

Therefore, the problem of identifying creative talent in this area has been an important one for those who wish to select engineers already in the field as well as for educators whose interests lie in the development of future designers.

The problem for the present study, then, is to develop instruments with which to identify creative machine design ability. The investigation is primarily concerned with making this identification at the industrial level, but an attempt to predict creative ability at the time individuals begin specialized engineering courses in college (usually the middle of the third year) is also being made.

Previous investigations in this area have been very limited. Some progress has been made by Guilford (2), in the factoring of tests assumed to measure reasoning ability. However, no validation of these tests has been undertaken to date. Until such validation is done on a sample of machine designers, nothing can be inferred about the extent to which creativity in machine design may be predicted from these tests.

Perhaps the most elaborate system for selection of prospective creative designers used in industry is the technique employed by the General Electric Corporation (5). A series of tests, interviews and job assignments, as well as college grades are used to evaluate candidates for G.E.'s Creative Engineering Program. Executives in charge of selection for this program seem well satisfied with the system, but again, the lack of validity data makes it difficult to obtain an objective evaluation of the instruments used.

One fact which emerges from the review of the available literature in the area is that the definition of creative ability differs somewhat among different investigators. Also, "creativity", in the technical sense, is not always the same as "creativity" used by the layman. Therefore, in order to avoid confusion of terms, the following definitions for creative and non-creative designers were adopted for purposes of this study.

> Creative designers: Persons who have demonstrated the ability to comprehend the nature of a design problem, and to produce a novel, ingenious, or original solution in the form of a total, functional, and practical mechanism. Creativity, in this sense, does not

necessarily involve the conception of an entirely new principle, but does involve the combination of existing principles or mechanisms in such a way as to produce a new and unique solution to a previously unsolved problem.

Non-creative designers: Persons whose major function is to work out the details of a design; that is, the engineers who do not produce original ideas, but who work out the routine problems of what materials to use, and who smooth out the design according to established procedures.

To obtain more suggestions about the types of tests which might be useful in predicting creative ability, persons in charge of design departments in which creative individuals were employed were asked for their opinions concerning the characteristics of creative engineers. On the basis of these suggestions and the previous studies in this area, four major assumptions were adopted as guiding principles for the construction of testing instruments.

First, it was assumed that in order to maintain a position as a machine designer, either creative or non-creative, an individual must possess a certain minimum amount of general intellectual ability. It was further assumed that this minimum was considerably higher than what is generally termed "average intelligence". For this reason, any attempt to discriminate between creative and non-creative designers on the basis of scores on standard intelligence tests would probably not prove fruitful.

A second major assumption was that a more accurate measurement of creativity could be obtained by evaluating something which an individual has produced instead of attempting to score judgments which he has made about something produced by another. To this end, completion type rather than recognition type tests were constructed, The third assumption made was that differences in age, experience, and formal training, apart from differences in creative ability, might produce differences in test scores between the criterion groups. To prevent the tests from becoming mere measures of time and training, the criterion groups were matched as nearly as possible with respect to these three variables, and analyses were performed to determine how well the matching had been done.

Finally, since the "specificity" or "generality" of creativity is unknown, it was assumed that the trait is specific to a given area. This is not to say that only engineers can be creative, but rather, that creativity in engineering is different from creativity in other fields of endeavor. On the basis of this assumption, problems dealing with machine design rather than with "general creativity" were developed.

II. METHOD

A. Tests Constructed

With the aforementioned assumptions as points of reference, six special ability tests and a personality test were constructed. All special ability tests were administered to college students to obtain an estimate of item difficulty and time necessary for administration. Items which proved extremely difficult at the college level were discarded, and two experimental test batteries were assembled from the remaining items. Battery A consisted of three special ability tests and a personality test. These were:

Mechanical Ingenuity, Power Source Apparatus Test

Three Dimensional Space Relations Test

Figure Matrices Test

Personal Inventory

Battery B was composed of the following special ability tests:

Mechanical Ingenuity, Design A Machine Test

Number Series Test

Unstructured Test for Creativity in Machine Design

A type of biographical information blank, the Personal History Form, was also included with both test batteries.

B. Selection of Item Analysis Sample

Industrial firms were contacted by mail and asked if they had groups of creative and non-creative engineers in their employment. The previously stated definitions of creative and non-creative engineers were included in the letters.

If such individuals were employed by these companies, and if the companies were willing to allow their engineers to be tested, they were instructed to have an executive in charge of the engineers (usually the chief engineer) identify those persons who could be considered creative according to the definition, and to select a like number of non-creative engineers, matching the two groups as nearly as possible with respect to age, amount of education, and amount of experience in the field. Subsequent t tests between the criterion groups on these variables yielded no differences which were significant at the 5% level of confidence.

Sixteen companies, contributing a total of 136 engineers, agreed to cooperate in the study. During a two-month period, Battery A was administered to nine of the participating companies, contributing a total of 70 men, and Battery B was administered to the seven remaining companies, providing a total of 66 men. The types of firms tested with Battery A included a company producing metal containers, a printing press corporation, a firm producing washing machines, a hydraulic appliance company, a machine tool plant, a firm which manufactures controls for machine tools, and two tractor manufacturing companies. Battery B was administered to engineers at an iron works, a brake and clutch company, a hydraulic appliance firm, a company manufacturing electric power line equipment, a tractor works, a machine tool plant, and a company which makes parts for jet engines.

C. Scoring

Scoring on the Three Dimensional Space Relations Test (3-D), the Figure Matrices Test (FM), and the Number Series Test (NS) consisted simply of marking items as correct or incorrect on the basis of a previously constructed scoring key. In addition, a part score was obtained for the items in the 3-D test. Since the answers involved the correct identification of three sides of a cube, each side was scored as correct or incorrect. The nature of the remaining tests was such that no scoring key could be devised beforehand. Rather, a list of possible scoring methods was compiled for each of these tests.

Items in the Mechanical Ingenuity, Power Source Apparatus Test (PSA), and the Mechanical Ingenuity, Design a Machine Test (DM) were examined for total number of solutions given. In addition, the solutions given were classified as workable or not workable by two independent judges. For the PSA test, the per cent of agreement between judges on solutions for individual items ranged from 66% to 87%, with a mean of 70%. For the DM test, interjudge agreement ranged from 69% to 93% for individual items, with a mean of 84%.

The Unstructured Test, (Unstr), was scored for total number of objects identified, number of responses per minute, number of responses involving motion of the objects seen, average number of line segments used in each response, number of machines identified, and per cent of machine responses

Items in the Personal Inventory (PI) were of the paired statement forced-choice type, the scoring of which could not be predetermined.

Instead, an analysis was performed, as described below. For the items that discriminated significantly between the criterion groups, a scoring key was determined, such that for each item, the response more characteristic of the creative group was identified as the "correct" response.

III. ANALYSIS OF THE DATA

A. Item Discrimination

As in scoring, the nature of the items as well as the number of available methods of analysis dictated the procedure for determining how well each item differentiated between the creative and the non-creative groups. For all tests of item discrimination, subjects from all companies were pooled, and tested as a single sample,

Table 1 shows the tests, scored as indicated, which were analysed by means of the H test (4), a test of the significance of the difference between ranks, to determine if a significant difference existed between the creative and non-creative groups.

Table 1

Tests and Scoring Methods:
Items Analysed by Means of the H Test

Test	Scoring Method
PSA	Total number of solutions
PSA	Number of workable solutions
DM	Total number of solutions
DM	Number of workable solutions
Unstr.	Total number of responses
Unstr.	Number of line segments per response

The tests for which contingency tables were set up to test for differences between the criterion groups are shown in Table 2. Finally, the part score and the whole score for the 3-D test items were used as predictors in discriminate functions in an attempt to predict the creative, non-creative criterion.

Table 2

Tests Analysed by Means of Contingency Tables

Tests	Scoring Method	# Rows X # Columns	Border Classifications
Unstr.	Motion Responses	2 X 2	creative, non-creative motion, non-motion
Unstr.	Rate of Responding	2 X 4	creative, non-creative minute 1, minute 2, minute 3, minute 4
PI		2 X 2	creative, non-creative right, left
FM		2 X 2	creative, non-creative right, wrong
NS		2 X 2	creative, non-creative right, wrong
3-D		2 X 2	creative, non-creative three sides correct, less than three sides correct

B. Tests for Company Effects

In addition to ascertaining whether or not the test items discriminated between creative and non-creative engineers, each item was analysed to determine whether responses differed significantly among the various companies tested. Two company effects were tested.

1. Company effect A. For this test, the following question was asked about each item: Are there companies that are contributing more or fewer correct responses (or other response measure, depending upon the scoring method used) than would be expected from the number of people in that company, regardless of the classification of these people with

respect to the criterion? To answer this question, a goodness of fit Chi square was obtained for each item, with the number of responses by each company as contributing elements.

2. Company affect B. For those items which differentiated between groups when individuals from all companies were lumped together, the following question was asked: Is the proportion of correct responses (or other response measure, depending upon the scoring method used) made by the creative individuals independent of classification by company? The answer to this question was provided by obtaining j X 2 Chi squares with the proportion of correct responses by criterion groups for each company as contributing elements. Snedecor's computational method (6) was employed to obtain these Chi squares.

C. Reliability

Odd-even correlations, corrected by the Spearman-Brown formula for a test twice the length of the halves, were obtained on all tests which were to be included in the revised battery. The information from the Personal History Form was such that it was impossible to compute equivalent halves reliability coefficients.

D. Item Difficulty

The percentage of subjects answering correctly was computed for all items. In the case of a scoring method involving number of responses as the measure used, the mean number of responses per item served as the index of difficulty. In such cases nothing could be inferred about the absolute difficulty of the item, but an estimate of the difficulty of a given item as compared to another item could be made.

E. Intercorrelations

For purposes of determining whether or not additional items should be included in the revised battery, an estimate of the degree of association between items was obtained. The procedure of intercorrelating every item with every other item was not followed for several reasons. Since all items had been selected on the basis of their relationship with the criterion, any measure of association between items would be spuriously high. Further, items were selected from two different batteries of tests which were administered to different samples of subjects. It was therefore impossible to intercorrelate items which were not common to a given battery. Finally, no estimate of item reliability could be obtained from the available data, but it might be expected that individual item reliabilities would be fairly low. With low item reliabilities limiting the magnitude of the correlations between items, an accurate estimate of inter-item association would have been very difficult if not impossible to obtain.

Considering the above restrictions, estimates of association were obtained among items on individual tests of the same battery. No correlations between items on different tests were attempted. However, total scores based on the number of discriminating items answered correctly were intercorrelated for tests within a given battery.

Tests for which item intercorrelations were obtained were the PSA, DM, Unstr., and PI. (For the PI, intercorrelations were obtained only on items discriminating at the .10 level or better.) Total score intercorrelations were obtained between PSA and PI, and between Unstr. and DM.

Product-moment correlations were used in all cases except for the intercorrelations among items of the PI. Those items were such that the Phi coefficient was a more appropriate measure of association.

IV. RESULTS

A. Item Discrimination

The criterion for the selection of an item to be used in the revised battery was that the null hypothesis of no difference between the means of the creative and non-creative populations be rejected at the .20 level of confidence or better. In addition, for each test, the CR described by Brozek and Tiede (1) was used to determine whether the number of items in that test which were significant at the .20 confidence level or better was greater than could have been expected by chance alone. Table 3 shows the number of items which were chosen from each test for further validation, according to the above criteria.

Table 3

Number of Items, by Tests and Scoring Methods, Selected for Validation

Test	Scoring Method	No. of Items	CR	p
PSA	Total no. of sol's	7	2.5	.0062
PSA	No. workable sol's	6	2.0	.0227
PI	----	50	3.70	.0001
Unstr.	No. of responses	6	2.06	.0197
DM	Total no. of sol's	4	1.65	.0495
DM	No. workable sol's	2	1.11	.1357

In addition, some ten areas from the Personal History Form were found to discriminate between the criterion groups, and have been included in the revised battery.

B. Company Effects

Tests for company effects showed that the items which had been chosed from the PSA and DM tests were free from both company effect A and company effect B. Two items of the PI, and three items from the Unstr. test were found to have a significant company effect A. All items except one from the Personal History Form were free from company effect B.

Since company effect A is primarily concerned with the relationship between companies, without concern for the classification of engineers within those companies, it would seem that a significant company effect A might be important in determining the norms for a given test or test item for a specific company. However, if all subjects in some companies had higher scores than all subjects in other companies, this would not imply that a given item was discriminating between the criterion groups in one company to a greater extent than in another company. For this reason, no item was eliminated merely because it was found to have a significant company effect A.

On the other hand, a significant company effect B would indicate that the proportion of correct answers given by a particular criterion group was dependent upon the company from which this group was taken. Therefore, it would be inferred that an item with a significant company effect B was not discriminating equally well in all companies. If this were the case, it would be justifiable to include such an item if the final battery were to be used only in those companies where the item discriminated well between the criterion groups. Since this was not to be the purpose of the final battery, the one item which had a significant company effect B was excluded.

C. Reliability

The reliability coefficients obtained are shown in Table 4.

Table 4

Reliabilities of Revised Tests

Test	r	Corrected r
PI	.62	.77
PSA	.78	.88
DM	.75	.86
Unstr.	.89	.94

As in the case of item intercorrelations, these r's may be somewhat spuriously high, since they were obtained using only those items which discriminated between the criterion groups. However, it would appear that these correlations are high enough to justify further testing without the construction of new items merely to increase test reliability.

D. Item Difficulty

The item difficulty indices generally ranged around .50, with the exceptions of items at the beginning and end of the 3-D and FM tests. Items at the ends of these two tests appeared extremely difficult, because most of the subjects did not have an opportunity to attempt them in the time allowed. Even though the amount of information was limited on such items, they were not included in the revised battery.

since with both tests, a large number of items which had been attempted by nearly all Ss did not differentiate between the criterion groups.

E. Intercorrelations

Phi coefficients between discriminating items on the PI are shown in Table 5. As previously mentioned, the item intercorrelations for the PI included only items discriminating at the .10 level. The corrected reliability coefficient for the PI, including those items discriminating at the .20 level, was .77, which would suggest that the item reliabilities are quite low. This is not uncommon for personality items. It would seem reasonable that those items discriminating at less than the .10 level may have done so at least partially because of lower item reliability. If this were the case, it would be expected that item intercorrelations based on these items would be spuriously low, and so low as to obviate any inferences from them, concerning the construction of new items.

From Table 5 it is apparent that the item intercorrelations for the PI, even among the items discriminating at the .10 level or better, were quite low. If items on the PI are measuring several distinct personality traits associated with creativity, the number of items which measure each trait was extremely small on the experimental test, and should be increased on the revised test. However, with the information which is available at present, it is not possible to say, with any degree of certainty, that unique factors are being measured by items on this test. In view of these facts, a compromise solution was reached.

Table 5

Item Intercorrelations: Personality Inventory
(Items Significant at the .10 Level or Better)

Item	38	42	47	59	61	67	68	71	98	101	105	109	113	116	121	124	129	148	160	179	192	Item
1	.09	.08	.00	.00	.32	.05	.13	.44	-.12	.20	.11	.03	.08	.41	.34	-.14	.22	-.08	.34	.08	.34	1
4	.09	-.07	.12	.10	.16	.38	.64	.17	.24	-.12	.00	.49	-.04	.13	.18	.16	.20	.01	.45	.36	.26	4
8	.45	.16	.44	.21	.30	.12	.24	.34	.00	.05	.00	.34	.28	.41	.24	.17	.21	.28	.34	-.14	.04	8
21	-.38	.05	-.03	.13	-.38	.12	-.14	.00	.65	.12	.46	.12	-.02	-.20	-.23	-.12	.23	.44	.14	-.02	-.24	21
23	.04	-.16	-.05	-.39	.28	-.07	-.04	.28	.07	.35	.18	.17	.11	.34	.30	.22	.25	.09	.06	.22	-.04	23
31	.30	.25	.30	.23	.30	-.08	.03	.07	-.19	.25	.07	.04	-.02	.46	-.04	.08	.23	-.16	-.35	-.12	.22	31
38		.21	.59	.00	.55	.38	.55	.06	-.18	-.03	-.20	.60	.16	.51	.07	.56	.10	.19	.07	.06	.26	38
42			.18	.20	.08	.32	.05	.38	.23	.10	.10	.07	.06	.31	.04	.06	.30	.14	.25	-.24	.34	42
47				.10	.42	.24	.58	.14	-.15	.53	.10	.32	.15	.43	.23	.35	.10	.03	.39	.05	.20	47
59					.07	.05	.10	.00	.40	.20	-.14	.27	-.04	.24	-.04	-.04	.24	.36	.20	-.04	.51	59
61						.44	.11	.04	.05	.15	-.17	.39	.13	.66	.52	.44	-.03	-.10	.41	.24	.11	61
67							.52	.07	.18	.04	-.19	.52	.22	.26	.27	.51	.34	.29	.24	.22	.10	67
68								.34	-.02	.18	.00	.49	.04	.28	.19	.14	.20	-.04	.14	.04	.24	68
71									-.04	.56	.37	.10	.22	.32	.45	-.03	.37	.00	.10	-.03	.22	71
98										.15	.36	.36	-.07	.16	-.11	.04	.29	.80	.30	.15	.09	98
101											.30	.01	.25	.33	.21	.05	.30	-.03	.09	.05	.20	101
105												.08	.36	.07	-.08	-.07	-.04	-.07	-.10	-.28	.00	105
109													.19	.52	.28	.54	.62	.32	.60	.42	.49	109
113														.07	.17	.11	.18	-.18	.14	-.32	.04	113
116															.62	.70	.24	.20	.18	.18	.37	116
121																.29	.08	.23	.07	.29	.30	121
124																	.07	-.08	.35	.00	.04	124
129																		.46	.20	.28	.10	129
148																			.14	.52	.24	148
160																				.24	.33	160
179																					.35	179

Item	4	8	21	23	31
1	.29	.18	-.03	.69	.17
4		-.04	.08	.35	-.26
8			-.28	.04	.21
21				-.18	-.20
23					.32

Items which discriminated at the .20 level were included in the revised test, but no new items were constructed.

Intercorrelations among items on PSA are shown in Table 6. These coefficients suggest that most of the items on this test are at least moderately associated. Exceptions are items 9 and 10 which appear to be somewhat unique. Since a large majority of the original items discriminated, and since the estimate of total score reliability was fairly high, no new items were constructed for this test. In the event that items 9 and 10 do not correlate with items in Battery B, an attempt will be made to construct similar items.

Table 7 shows the associations between the discriminating items from DM. While the correlations obtained are somewhat lower than those for PSA, no additional DM items have been constructed to date, since the structure and content of the DM items closely resembles that of PSA. It is expected that items from these two tests will be at least moderately correlated when it is possible to make comparisons between them. If such is not the case, additional DM items will be built.

Inter-item correlations for the Unstr. test are shown in Table 8. All discriminating items on this test were fairly highly intercorrelated, and total score reliability was the highest obtained for any test. From these results no new item construction seemed warranted for this test.

For total score intercorrelations, all items which discriminated at the .20 level or better were used in each case. These correlations are shown in Table 9. The amount of association between the tests in Battery A was slightly greater than in Battery B. However, neither of

Table 6

Item Intercorrelations:
Power Source Apparatus[a]

Item	1W	2W	3W	4T	5T	5W	6W	7T	9T	9W	10T
1T	.65	.43	.33	.48	.44	.36	.33	.63	.39	.23	.32
1W		.53	.43	.51	.42	.48	.55	.61	.39	.26	.30
2W			.44	.45	.45	.38	.49	.57	.50	.38	.31
3W				.32	.34	.39	.42	.47	.35	.28	.36
4T					.56	.49	.25	.49	.40	.22	.26
5T						.57	.39	.51	.41	.35	.27
5W							.38	.39	.24	.19	.31
6W								.47	.37	.27	.25
7T									.44	.31	.31
9T										.86	.20
9W											.15

[a] T signifies item scored total number of solutions. W signifies item scored number of workable solutions.

Table 7

Item Intercorrelations:
Design a Machine Test[a]

Item	3T	4T	7T	1W	4W
1T	.31	.26	.15	.58	.15
3T		.37	.46	.36	.34
4T			.33	.32	.56
7T				.23	.12
1W					.17

[a] T signifies item scored total number of solutions. W signifies item scored number of workable solutions.

Table 8

Item Intercorrelations:
Unstructured Test

Item	3	5	7	8	10
2	.72	.67	.64	.67	.60
3		.67	.70	.70	.65
5			.77	.73	.64
7				.78	.61
8					.71

Table 9

Total Score Intercorrelations
within Batteries

Test	PSA	Unstr.
PI	.455	
DM		.370

these correlations was high enough to justify the combination or deletion of tests from the revised battery.

In general it might be said that the results of the measures of association obtained among discriminating items on the several tests were somewhat restricted, both by item unreliability, and by the fact that all items were correlated with the criterion. Therefore, the conclusions drawn from these results must be tentative at best, and any decisions as to the construction of additional items should probably not be made on the basis of these conclusions alone. Rather, these intercorrelations were viewed as very rough approximations of the relationships which exist, and further item construction was postponed until more information about present items could be collected.

V. FUTURE PLANS

As has already been mentioned, arrangements have been made to test a sample of mechanical engineering upperclassmen in three advanced engineering courses. Subjects will be given all of the revised tests, and item intercorrelations will be obtained. In addition, two of the classes to be tested will be sections of courses dealing with machine and machine element design. For subjects in these sections, an intermediate criterion, consisting of classification by the section instructor, will be available in the near future. With this information, an estimate of the predictive efficiency of the revised tests at the college level may be obtained.

Once intercorrelations among all items are available, the final test battery will be assembled. Additional items will be constructed if unique factors appear, or some consolidation may be made if intercorrelations are high. With this final battery, a validation study will be undertaken. Industries are now being contacted to obtain the subjects necessary for this study.

Further work will be contingent upon the validity of the revised battery. If a highly valid battery of tests is obtained, the next step would logically be the standardization of these instruments on a large sample of engineers from many different industries, and in various geographical locations. If the validity of the battery is too low for satisfactory prediction, new tests may be constructed in the hope that other more fruitful aptitudes and abilities may be measured.

VI. LITERATURE CITED

1. Brozek, J. and Tiede, K. Reliable and questionable significance in a series of statistical tests. *Psychol. Bull.*, 1952, 49, 339-341.

2. Green, Russel F., Guilford, J. P, Christense, Paul R. and Comrey, Andrew L. A factor-analytic study of reasoning abilities. Psychometrika, 1953, 18 No. 2, 135-160.

3. Johnson, Palmer O. *Statistical methods in research.* N.Y., Prentice-Hall, Inc. 1949.

4. Kruskal, W. H. and Wallis, W. A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assn.*, 1952, 47, 583-621.

5. Schmall, W. A. The General Electric Company. Personal communication. 1952.

6. Snedecor, George W. *Statistical Methods* 4th ed. Iowa, The Iowa State College Press, 1946.