GRE

GRADUATE RECORD EXAMINATIONS

MEASURES FOR THE STUDY OF CREATIVITY IN SCIENTIFIC PROBLEM-SOLVING

Norman Frederiksen
and
William C. Ward

May 1978

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

MEASURES FOR THE STUDY OF CREATIVITY IN SCIENTIFIC PROBLEM-SOLVING

Norman Frederiksen
and
William C. Ward

May 1978

Reprinted from *Applied Psychological Measurement*
Vol. 2, No. 1, Winter 1978 pp. 1-24

MEASURES FOR THE STUDY OF CREATIVITY IN SCIENTIFIC PROBLEM-SOLVING

Norman Frederiksen
and
William C. Ward

A set of Tests of Scientific Thinking were developed for possible use as criterion measures in research on creativity. Scores on the tests describe both quality and quantity of ideas produced in formulating hypotheses, evaluating proposals, solving methodological problems, and devising methods for measuring constructs. The tests were administered to 3,500 candidates for admission to graduate school in psychology, using an item-sampling procedure. Reliabilities based on 45-minute tests were adequate for research purposes. Correlations with GRE scores were low, especially for scores based on number of ideas. Follow-up questionnaires were sent to students asking for information about graduate school attendance, grades, accomplishments during the first year of graduate study, and self-appraisals of professional skills. Scores based on quantity (number of responses, number of unusual responses, and number of unusual responses that were also of high quality) were significantly related to self-appraisals and to reports of such professional accomplishments as collaborating in research, publishing scientific papers, and designing and maintaining research apparatus. The quantity scores also were related to indices reflecting the quality of the department attended and to conventional evaluations of student performance. GRE scores were better at predicting these indices of quality but poorer as predictors oı accomplishments and self-appraisals.

Research on creativity is handicapped by a lack of adequate criterion measures. One is forced to rely either on subjective judgments or various "creativity tests." However, subjective judgments are often biased and untrustworthy, while the "creativity tests" have items which are too trivial to measure such a complex characteristic, although they may measure abilities related to creative performance. Much of the difficulty arises because creativity is not a distinct characteristic found in isolation but an aspect of performance exhibited in dealing with complex, challenging tasks. If a solution to such a problem is seen as particularly effective, as well as unusual or imaginative, it is judged to be creative.

The purpose of this investigation was to develop a set of tests that may serve, at least provisionally, as criterion measures in subsequent investigations of creative thinking. Such tests should be sufficiently complex to represent problem situations confronted in reality, yet sufficiently simple to allow for the control and standardization required of well-designed psychometric procedures. An attempt to measure creativity in the area of behavioral science (more specifically, psychology) began the investigation with the idea that, if successful, the methods might later be extended to other areas. The scope of the present work was to (1) construct a set of tests to elicit creative performance and (2) assess the acceptability of the tests from the standpoint of their psychometric properties (e.g., reliability, difficulty, interrcorrelations) and, to a limited degree, their validity.

The Tests of Scientific Thinking

Rationale for Test Development

Two major approaches to the study of creativity are typified by the work of Guilford and of MacKinnon. On the one hand, the Guilford approach is highly analytical involving the measurement of hypothesized independent factors within the "structure of intellect" model (Guilford, 1964, 1967). His tests of divergent production, along with their conceptual descendants, are often employed as criteria of creative performance. The MacKinnon approach (e.g., MacKinnon, 1962), on the other hand, involves the selection of individuals recognized as outstanding in a particular field, and attempts through intensive assessment to discover ways in which they differ from less outstanding practitioners of the same profession.

From the standpoint of the present objective neither approach was satisfactory. Available divergent thinking instruments, such as Guilford's, were too simple to provide convincing criterion measures, while methods such as MacKinnon's offered too little control over possible sources of confounding. Between the extremes of real-life criteria and simple analytic procedures, a middle ground was needed. It was proposed that situational tests be used which would be measures of performance based on situations resembling the problems which scientists encounter in their work. They would be constructed so that outcomes could be evaluated objectively and solutions could be compared directly.

The construction of situational tests would be facilitated by the availability of job analyses that describe the domain of situations and tasks comprising the work role in question; the tests could then be constructed by sampling from that domain (Guion, 1977). However, no such job analyses exist. For the purposes of this investigation, the most useful description of scientists' work is that given by Flanagan (1949), who had studied critical incidents in the activities of research personnel employed in 20 research laboratories. From these he derived a list of eight basic tasks in the role of the researcher, of which three seem particularly relevant to scientific thinking: (1) formulating problems and hypotheses; (2) planning and designing the investigation; and (3) interpreting research results.

Description of the Tests of Scientific Thinking (TST)

Prototype items for situational tests of various kinds were developed and pretested informally. Four types were eventually chosen, which appeared to represent important aspects of the job of a scientist, as revealed by the Flanagan critical incidents study. In addition it appeared feasible to develop appropriate items to measure the constructs implied by the four titles.

1. *Formulating Hypotheses (FH).* Each problem consists of a brief description of a research study, a graph or table showing the principal results, and a statement of the major finding. The task is to write hypotheses that might explain, or help to explain, the finding. The subject is asked to write not only the hypothesis he/she thinks is most likely to be correct, but also other hypotheses that should be considered in interpreting the data or in planning another investigation.
2. *Evaluating Proposals (EP).* The subject is asked to suppose that he/she is teaching a senior course in design and methodology, and that as a class exercise has asked the students to write brief proposals of research studies. Several proposals presumed to have been written by the students are presented as the test items. The subject's task is to write critical comments for each student regarding the design, methodology, or theoretical position of the proposal.
3. *Solving Methodological Problems (SMP).* Each problem is a brief statement of a methodological problem encountered by a graduate student or psychologist in planning a research study. The subject's task is to write suggested solutions to the methodological problem.
4. *Measuring Constructs (MC).* Each problem consists of a name and definition of a psychological construct (e.g., conservatism, bigotry, leadership). The task is to suggest methods for eliciting the behavior so that it can be observed and measured without resorting to ratings or self-report methods.

For each test the directions are followed by a sample problem and a sample answer sheet filled out by a hypothetical student. The sample responses were chosen to suggest the kinds of thinking the particular test was intended to elicit. (Figure 1 shows a sample item for Formulating Hypotheses.) In the case of all tests except EP, the subject is asked to mark the answer he/she considers to be the best.

Responses to these job-sample tests are complex. If the tests were at all successful, then they would not only reflect originality and flexibility,

Figure 1
Sample Problems and Answers

FORMULATING HYPOTHESES

Birth Weight and IQ

The IQ scores of 822 children aged 8 and 10 years were studied in relation to their birth weights. IQ was measured on the Wechsler Intelligence Scale for Children, while birth weights were obtained from hospital records. Results were as follows:

Relation Between Birth Weight and IQ

Birth Weight (Grams)	Number of Children	Mean IQ
1000-1499	46	84.8
1500-1999	69	88.3
2000-2499	302	91.2
2500-3000	405	95.9

Finding: Children who weighted more at birth tended to have higher IQs in middle childhood.

Suggested Hypotheses

Heavier babies came from higher SES families where mothers are more motivated regarding child care.	☒
Babies of lower weight included many premature births.	□
Mother's expectations regarding size of baby determined the level of intellectual development — ie. large babies were encouraged to be aggressive and adventuresome.	□
Heavier weight is due to larger and heavier brain.	□
The same factors of the mother's diet that determine heavier birth weight also determine higher intelligence potential.	□
	□

Mark the hypothesis you think is best by putting an X in the box at its right.

but also would be influenced by other cognitive abilities such as reasoning and by knowledge of behavioral science. They might, in addition, reflect temperamental characteristics that might, for example, result in self-censorship of ideas. It is this complexity (hopefully approaching that of real-life performance) that will make the responses useful for later studies of creative processes.

Scores and Scoring Methods

Two methods of scoring for response quality were investigated in an earlier study (Frederiksen & Evans, 1974) using a version of Formulating Hypotheses designed for the general undergraduate population. In the first, the scorer simply rated each response on a nine-point scale. In the second, the scorer classified actual responses to the item by assigning each response to one of a list of categories. The categories on the list were rated for quality, and a scale value was assigned to each. Thus, it was possible to have a computer assign the appropriate scale value to each response given by a subject.

For an N of almost 400, the correlation between quality scores derived by these two methods was .98, when corrected for unreliability. It was decided to use only the category scoring in the present study, since the judgment required for this method was less difficult than that required for the rating method and was, therefore, more economical.

The scoring method that evolved produced six scores for each test item. Three of these are directly concerned with the quality of the responses, two represent counts of responses, and one involves a combination of these two aspects of performance. The six scores are:

1. *Mean Quality:* The average of the quality values assigned to each response to an item.
2. *Highest Quality:* The quality value of the response with the highest scale value according to the scoring system. This score reflects the subject's best performance and gives credit for occasional excellence without being affected by a large volume of more pedestrian attempts.
3. *Best Quality:* The quality value of the response to an item, indicated by the examinee to be his/her "best." (This score was obtained for each test except Evaluating Proposals, where it was felt that the miscellany of bases on which a proposal could be criticized made such a comparison questionable.) This score may reflect the subject's ability to evaluate his/her own ideas.
4. *Number of Responses:* The total number of scorable nonduplicate ideas. This score gives almost no weight to the quality of the idea; however, a response had to meet a minimum criterion of comprehensibility and could not be a duplicate of another response. Since high productivity seems to be characteristic of individuals who produce excellent products, this score is of interest as an additional characteristic of productive thinking.
5. *Number of Unusual Responses:* The number of responses to an item which meet a criterion of infrequency in the distribution of responses. Generally, an unusual response is one which accounts for no more than 5% of all responses given by the examinees.
6. *Number of Unusual-High Quality Responses:* The number of responses to an item which meet criteria for both infrequency and quality. The quality criterion was that the category into which the response fell was rated in the top third of all categories for the item.

Number of Unusual Responses and Number of Unusual-High Quality Responses are subsets of the total number of responses. Each provides a method for viewing task performance which matches a frequently used definition of creative production.

These scores do not constitute six independent dimensions of task performance; there are both logical and experimental interdependencies among subsets of the scores. As a set, however, they appear to provide a fair representation of the major ways in which responses differ.

Development of the set of categories for each problem involved the coordinated efforts of the two authors and two experienced research assistants. Working independently, two of the four constructed a trial classification based on 50 protocols. After agreeing upon the list, the remaining two (working independently) attempted to assign all responses from the same 50 protocols to these categories. The discussion of disagreements and ambiguities led to further changes, until a consensus was reached. Two new scorers were then given the list of categories, along with a new set of protocols; their difficulties led to a final revision. This sequence of activities led to establishment of sets of categories which have proved able to accommodate approximately 95% of all responses obtained, with relatively few scoring problems.

In order to derive scores from the categorized protocols, a table of scale values representing the quality score to be assigned to each category was required. The two authors and two research assistants independently ranked the categories for each problem in order of quality, which was defined in terms of the intructions given examinees for the particular test. For example, for the Formulating Hypotheses items, the highest ranks were given for ideas judged most likely either to provide a correct explanation for the finding or to deserve serious consideration as competing hypotheses. In all cases, rankings depended primarily on judgment and general knowledge of psychological principles rather than on knowledge of specific results of investigations; none of the problems were such that a single neat solution could be derived from previous work.

Agreement in rankings among the four judges was, in general, very high. Alpha coefficients for individual items ranged from .76 to .98, with a median value of .92.

Administration and Scoring of Tests

The time (25 minutes) normally reserved for pretesting new items for the Graduate Record Examination (GRE) Advanced Psychology Test was used for the tryout of the experimental tests. Through an item-sampling method, it was possible to try out four six-item tests in one national administration of the GRE. Measures useful in evaluating construct and discriminant validity were available for these candidates: the GRE Aptitude Tests (V and Q) and the Advanced Psychology Test, with its two subtest scores. The sample for this study constituted a cohort of students planning to embark on a career in psychology. The possibility was thus introduced for follow-up studies of the predictive validity of the tests for both graduate school and later performance.

Examinees were informed that the experimental tests would not affect their GRE test scores. They were asked to sign a consent statement giving permission for use of their GRE scores in the analysis. In addition, they were asked to give permission for future contact for participation in a possible follow-up study.

Item-Sampling Method

The rationale for item sampling (Lord & Novick, 1968, Chap. 11) requires that subsets of items be administered to subgroups of examinees. If these subgroups are randomly chosen and are sufficiently large, variances and covariances of items obtained for each subgroup will provide good estimates of those that would have been obtained from the total group. Thus, item variance-covariance matrices may be assembled and treated as though all candidates had taken all items. From these matrices, unbiased estimates may be made of the reliabilities and intercorrelations of the six-item tests. By adding to the matrices the variances and covariances involving the various GRE scores and questionnaire items, the correlations of the four tests with these variables may also be estimated.

It was planned that 40% of the candidates (the "reliability sample") would each be given three items from a single test, while 60% (the "intercorrelation sample") would each be given two items from different tests. Since two tests were involved in the intercorrelation sample, two sets of instructions were read and two sample items were studied by the candidates. Both the formal and informal tryouts of the tests showed that seven or eight minutes per item was considered to be sufficient by most students; therefore, the time allowance was believed to be adequate. Although items were not separately timed, students were informed that they had 25 minutes to read the instructions, study the sample item (or items), and write their answers. It was assumed that time per item would be approximately the same for the reliability and the intercorrelation samples.

Two-item tests and three-item tests were assembled in such a way that all the combinations of two items from different tests in all the possible orders were present in equal numbers, as were all those of three items from one test. These experimental tests were then arranged in random order and placed in the Advanced Psychology Test booklets.

Description of Sample

The test administration took place in October 1973. The total number of candidates taking the Advanced Psychology Test within the United States in that administration was 4,394. Of these 3,586 provided complete data on the items they were given and were included in the analyses. About 800 cases were lost; half of these were refusals to participate, either overtly or through failure to sign the permission statement, while the remainder resulted from data that were incomplete or spoiled.

Several background information questions are routinely asked of candidates taking the GRE Advanced Psychology Test. Answers to these questions indicated that the typical candidate was a senior psychology major planning to attain a doctorate in psychology. More than two-thirds of the group had training in statistics, as well as in general and experimental psychology. Just over half planned a career in clinical psychology. In these respects those in the complete data sample differed little from the total group. Mean GRE scores were also very similar.

Background question responses were compared with those from other administrations of the Advanced Psychology Test. There were no appreciable differences between the complete data sample and a group constituting about 14,000 aspirants to graduate work in psychology. The sample was thus representative of individuals who wish to enter a career in psychology.

Scoring

Scoring was done by part-time, at-home workers, all of whom had bachelor's or master's level training in either psychology or a closely related discipline. Two scorers were assigned to each item; each completed all the protocols for that item before being trained to score another one. An assistant checked each set of 50 protocols when it was returned. The categorized protocols were then keypunched and stored on magnetic tape. A computer program was written to derive the six scores for each of the 24 items for the two scorers both separately and combined.

Scoring Reliability

Coefficient alphas (Cronbach, 1951) were computed for single items, based on categorizations by two independent scorers. The highest scoring reliabilities were those for the Number of Responses scores; the median over all four tests was .90. The three quality scores had similar, somewhat lower, reliabilities, with medians of .83, .80, and .80 for Highest, Mean, and Best Quality, respectively. The lowest median reliabilities were those for Unusual (.72) and for Unusual-High Quality (.69) responses, presumably since (by definition) number of responses contributing to these scores is small. Scorer reliability was high enough, therefore, to justify the use of all of the scores in further explorations of the psychometric properties of the tests.

Psychometric Properties of the Tests

Test Reliability

The reliabilities of the six scores for the four Tests of Scientific Thinking are shown in Table 1. These coefficients reflect both the agreement among scorers and the consistency in performance on the part of the subjects. They are estimated reliabilities for six-item tests; they were calculated under the assumption that each subsample of individuals, given a particular pair of items from one of the tests, provides an unbiased estimate of the result which would have been obtained had all individuals been given a six-item test. In the calculation of coefficients for these tests and of other estimated coefficients for six item tests (if individuals given the various subsets of items were perfectly comparable), the estimates derived would be exactly the same as those which would have been yielded by complete six-item tests.

Two estimates are given in each cell of the table. The first entry is a lower bound reliability estimate, Guttman's (1945) "Lambda 2"; this coefficient is similar to, but generally higher than, coefficient alpha (Koutsopoulos, 1964). The second entry may be thought of as an upper bound coefficient (Cronbach, 1951); it is an estimated parallel-form test-retest reliability, as-

Table 1
Reliabilities of Scores on the Tests of Scientific Thinking[a]

Test		Score					
		Best	Mean Quality	Highest	Number	Unusual	Unusual-High
FH	lower bound	.50	.62	.65	.67	.53	.46
	upper bound	.62	.71	.79	.78	.69	.59
EP	lower bound		.61	.51	.73	.55	.44
	upper bound		.79	.60	.81	.74	.58
SMP	lower bound	.30	.46	.35	.61	.42	.05
	upper bound	.51	.62	.48	.73	.60	.21
MC	lower bound	.68	.77	.71	.77	.36	.17
	upper bound	.80	.88	.85	.82	.46	.34

[a] *N*'s for the reliability sample were 359, 339, 309, and 340, for FH, EP, SMP, and MC, respectively. Each of these examinees provided scores on three items from one test, except that a number of examinees sometimes omitted marking a "Best" response. Effective *N*'s for Best response are 323, 248, and 290 for FH, SMP, and MC, respectively; examinees were not asked to designate a best response on EP.

suming two hypothetical forms which are maximally similar. The upper bound estimates may be used to provide conservative estimates of true-score correlations of the experimental tests both with each other and with other measures. For other purposes, lower bound estimates are more conservative.

The two sets of estimates differed considerably, with a median difference of .13; their pattern, however, was the same. SMP was, in general, the test with the lowest reliabilities, while MC had the highest. Across tests, the Number score tended to be most reliable and the most consistent. The Unusual-High reliabilities for SMP and MC were particularly low, but for the remaining two tests, this score was reliable enough to be useful.

Level of Performance

The means and standard deviations of the scores, based on the reliability sample, are shown in Table 2. Candidates had more ideas (almost four per item) for Evaluating Proposals than for any of the other tests. FH and SMP elicited approximately two and one-half ideas per item, while MC elicited slightly more. The number of responses classified as Unusual was roughly one-fourth of the total number of ideas per item, while the number classified as Unusual-High was generally less than one-third.

The mean number of response categories associated with the items (and therefore the highest possible rank for any of the quality scores) varied from test to test (see Table 2, footnote b). The mean Quality scores actually obtained were slightly above the midpoint of the scale for each test. Responses designated by the students as their "Best" were scored a point or two higher than the mean, on the average, while those which were "Highest" were approximately two points higher.

The numbers shown in Table 2 are based on the reliability sample, those students who took three items from one test. It had been assumed that the time available per item would be approximately the same for both samples, since those in the intercorrelation sample were required to study two sets of instructions and two sample items instead of one. It was, therefore, expected that the means and variances would be about the same.

A comparison of the means for the two groups revealed that for the three quality scores the dif-

Table 2
Means and Standard Deviations of Scores on the Tests of Scientific Thinking
(Based on Reliability Study Sample)[a]

Test		Score					
		Best[b]	Mean Quality	Highest	Number	Unusual	Unusual-High
FH	Mean	19.67	17.88	22.09	2.45	.61	.24
	SD	3.42	2.97	2.91	.57	.36	.22
EP	Mean	[c]	17.80	23.94	3.90	1.10	.18
	SD		1.94	1.84	.93	.49	.18
SMP	Mean	14.89	14.00	17.18	2.42	.52	.12
	SD	2.23	1.95	1.84	.57	.29	.11
MC	Mean	14.55	13.70	17.95	2.78	.77	.21
	SD	3.64	2.96	3.20	.79	.36	.17

[a] N's are approximately the same as those shown in Table 1.
[b] The highest possible values for the Quality scores vary from test to test. For FH, this value is 25.42; for EP, 27.00; for SMP, 22.29; and for MC, 25.04.
[c] Candidates were not asked to choose their best answer for EP.

ferences were, indeed, minimal. In no case did they differ by more than 4% (.3 standard deviations). For the three scores involving counts of items, however, the means were all higher for the intercorrelation sample; and the differences were sometimes quite large. Students taking only two items wrote (on the average) 15% more responses per item, gave 16% more Unusual Responses, and gave 15% more responses which were both Unusual ańd of High Quality.

These differences suggest that the time required to read the instructions and study the sample item for the second test was sufficiently brief to allow slightly more time for responding to the items. This extra time resulted in the production of more responses without appreciably influencing the quality of the ideas. Apparently, the best ideas tended not to be produced near the end of the allotted time; and the extra time resulted in additional answers of only average quality.

Correlations of Scores Within Each Test

Correlations among scores derived from any one of the tests can be estimated from a matrix of variances and covariances of the six scores for each of the six items. Data supplied by the reliability sample provided all the terms needed.

The coefficients were computed in two ways. The first made use of all relevant inter-item covariance terms, including scores obtained from the same item. This provided an estimate of the correlations which would have been obtained had the complete six-item tests been given to all students (without item sampling). A second procedure was developed in which correlations were estimated, using only covariance terms for scores obtained from different items. The coefficients estimated in this way were estimates of the intercorrelations that would have been obtained by giving each student a different set of six test items to yield each distinct score. This method eliminated experimental dependencies between scores and was useful in assessing the degree of relationship among the *abilities* represented by the several scores. Table 3 presents correlations computed according to the second method.

For each test the quality scores tended to form a cluster, although the tendency was less marked for SMP. The three count scores also tended to cluster in the cases of FH and EP. These groupings occurred despite the elimination of experimental dependencies among scores.

Table 3
Intercorrelations of Scores for Each Test after Eliminating Experimental Dependencies

Test		Score				
		Mean	Highest	Number	Unusual	Unusual-High
FH	Best	.50	.48	.20	.20	.28
	Mean		.52	.05	.03	.15
	Highest			.31	.13	.20
	Number				.40	.31
	Unusual					.35
EP	Best	--	--	--	--	--
	Mean		.44	-.03	-.20	.01
	Highest			.35	.10	.18
	Number				.56	.28
	Unusual					.26
SMP	Best	.43	.21	-.28	-.27	-.02
	Mean		.20	-.37	-.42	-.08
	Highest			-.00	-.25	-.06
	Number				.30	.01
	Unusual					-.14
MC	Best	.64	.60	.03	.09	.26
	Mean		.65	-.03	-.11	.26
	Highest			.24	.08	.29
	Number				.49	.20
	Unusual					.03

The correlations between the quality scores and the count scores tended to be low, although the pattern was not consistent. These coefficients were all negative for SMP. It might be assumed that for SMP, those who wrote responses of high quality (responses solving the problem) thought that additional answers would be irrelevant.

An interpretation of the intercorrelations computed by the other method would not be different from that given for Table 3, even though the intercorrelations were substantially higher. The magnitude of the coefficients indicates that the information provided by the six scores was not redundant.

Correlations of Scores from Different Tests

The intercorrelation sample was comprised of 60% of the candidates tested. The sample was intended to provide estimates of the correlations among scores from the four Tests of Scientific Thinking. One-sixth of this group was assigned to each of the possible pairs of tests. Within each such subgroup, each candidate was given one item from each of the two tests All possible item pairs were administered in both possible orders.

Computation of the correlation between any two scores requires use of the inter-item covariances and estimated total test variances for those scores. The covariances present no difficulty; however, the test variances must be estimated using data from the reliability sample, since that group alone provided the necessary covariance terms for pairs of items within a test. The study design, therefore, called for the use of estimated test variances from the reliability sample in the calculation of inter-test correlations.

For the quality scores, the comparability of samples, with regard to item means and variances, suggests that this procedure was justified. For the count scores, however, the comparisons suggest that an item presented in a two-item context should be treated as a longer test than the same item given in a three-item context. The ratio of means under the two sets of conditions may be used as an index of the amount of lengthening. Using Gulliksen's (1950) formula for the variance of a lengthened test, this ratio provided a basis for the necessary correction in the test variances estimated for the reliability sample.

The entries in the intercorrelation matrix that are of most interest are the correlations between corresponding scores from different tests. These correlations may indicate whether an individual's production of ideas or quality of ideas is uniform across different kinds of tasks, or whether the ability to produce solutions to problems is specific to the kind of problems posed. The estimated correlations are shown in Table 4.

If the coefficients within each column of the table were uniformly high, it would suggest that problem-solving performance was constant, regardless of the type of task required of the subject. Moreover, it would indicate that items from the four test types could be combined and only six scores used on one composite Test of Scientific Thinking. In fact, the correlations varied quite widely, not only within each column, but also within each row. It does not appear justifiable to combine analogous scores obtained from different tests, except in specific instances. Nor can one assume that a pair of tests will be comparable in eliciting similar behaviors over the whole spectrum of scores. There is little evidence of a generalized ability to produce ideas which are either numerous or good.

There is one "impossible" coefficient (1.03) in Table 4. Apparently, despite the correction for length, test variances for scores in the intercorrelation sample had still been underestimated. In an attempt to improve the estimates, a second computational procedure was devised in which one item was considered to constitute a unit-length test; the problem was seen as that of estimating the variance of such a test lengthened to six units. Gulliksen's formula was again applied. It was assumed that the reliability of each score was identical for the intercorrelation and reliability samples; otherwise, no information from the latter sample was used. When variances derived in this manner were used in computing correlations, the resulting coefficients did not differ appreciably from those in Table 4. There was a correlation of .98 between the two sets of coefficients over these 23 test-by-score combinations. Thus the *pattern* of coefficients is quite stable over two methods which differ substantially in the assumptions upon which they are based.

Configurational Analyses

Two types of configurational analyses were performed: The first employed the Guttman-Lingoes Smallest Space Analysis (Guttman, 1968); the second one used factor analysis. The basic input for both series was a 19 × 19 correlation matrix involving five scores from each of three tests and four scores from the remaining test. The Unusual-High scores were not included in these analyses, primarily because these scores were composites of both quantitative and qualitative aspects of test performance.

Experimental dependencies had been eliminated from the correlations among scores within each test used in the analysis. Correlations among scores across tests were those estimated using test variances which were obtained from the reliability sample and corrected for differences in test length. The one estimated correlation in the matrix which exceeded 1.0 was re-

Table 4
Correlations of Corresponding Scores from Different Tests

	Best	Mean Quality	Highest	Number	Unusual	Unusual- High
FH-EP	--	.53	.18	.68	.43	.22
FH-SMP	.37	.06	.11	1.03	.16	.35
FH-MC	.38	.24	.35	.48	.19	-.26
EP-SMP	--	.63	.76	.42	.04	-.23
EP-MC	--	.53	.60	.30	.04	.21
SMP-MC	.34	.17	.32	.59	.42	.41

placed by the product of the square roots of the reliabilities of the two scores involved; this is equivalent to assuming that the true score correlation between the two variables is 1.0.

Smallest Space Analysis

Since Smallest Space Analysis attempts to preserve only ordinal relations among the coefficients, it may produce a solution which is less complex and more defensible than that obtainable through factor analysis. Given the way in which the matrix was derived, the interval score assumption required for factor analysis is likely not to have been met.

Analyses in two-, three-, and four-vector space were performed. The two-vector solution (see Figure 2) provides the simplest and most interpretable picture of the data.

Vector 1 is clearly a contrast between the quality scores (all of which had negative coordinates on this axis) and the count scores (all of which had positive coordinates). Within the quality scores, there was a further ordering of the three scores derived from each test: the highest negative coordinates on Vector 1 were associated with Mean Quality, while the lowest tended to be associated with Highest Quality. For the count scores, the highest positive coordinates tended to go to Unusual scores rather than to scores for Number of Responses. The configuration suggests the use of Mean Quality and of Unusual scores to provide the maximal discrimination between quantitative and qualitative aspects of test performance.

The second vector in the Smallest Space solution provided several contrasts between the quality scores. The major distinction was that SMP and EP quality scores all had positive coordinates, and MC and FH quality scores all had negative coordinates. Furthermore, all three FH scores were given high negative coordinates, while MC scores were located more centrally. Thus, there were at least two, and possibly three, aspects to quality of response on the experimental tests.

Additional Smallest Space Analyses did not add any further suggestions of relations among the various scores. An increase in the dimensionality of the space produced no meaningful ordering along the third or fourth vector. Adding the Unusual-High Quality scores resulted in configurations virtually identical to those produced when the Unusual-High Quality scores were omitted. Finally, analyses in two-vector space for the quality scores alone and for the count scores alone merely repeated the ordering of scores found within the complete analysis.

Smallest Space Analysis thus revealed three aspects of the relations among the various scores. First, all the scores derived by counting responses were distinguished from all those derived from evaluations of quality. Second, the quality scores EP and SMP, were similar in location to one another and distinct from FH and MC. Third, the quality scores FH and MC, were somewhat distinguishable from one another.

If GRE Verbal and Quantitative scores and the two Advanced Psychology subscores were added to the matrix, the GRE scores formed a tight cluster. This reflects their relatively high intercorrelations (.43 to .67).

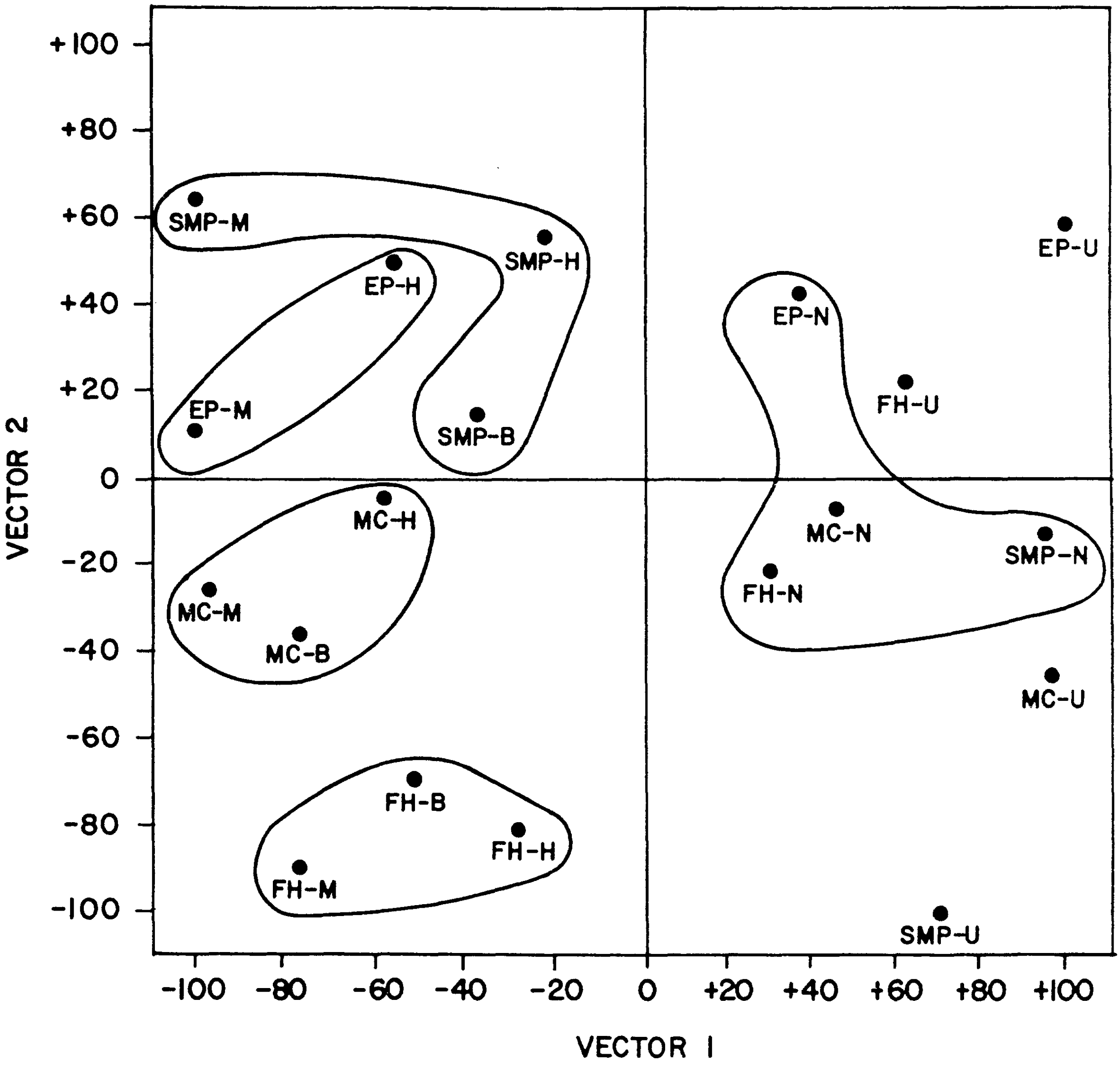
Factor Analysis

A series of factor analyses of the 19 × 19 correlational matrix was undertaken. Principal factor solutions for three, four, and five factors were obtained and were subjected to both varimax (orthogonal) and oblimin (oblique) rotations. In the initial analyses, iterated communalities for some scores exceeded 1.0 (which is another indication of inaccuracies attributable to estimations required by item sampling). In order to avoid this problem, the square of the largest off-diagonal element in each column was used, without iteration, as an estimate of communality.

The three-factor varimax solution was favored, both on the basis of the plot of magnitudes of successive roots and for reasons of interpretability. This solution is shown in Table 5.

The first factor is clearly a general count-score factor. With one exception, the Number and Unusual scores from all four tests had their highest loadings (all ⩾ .40) on this factor. Factor II appears to be a quality factor primarily for FH and MC, and Factor III is a quality factor based primarily on EP and SMP. There was, however, some overlap in the quality factors; one score each from EP and SMP had a substantial loading on Factor II, as well as loading on Factor III. In addition, one of the count scores—that for Unusual Responses on

Figure 2
Two-Vector Smallest Space Solution[a]



[a]Each score is designated by the abbreviation for the test (FH, EP, SMP, MC) followed by a letter. B, M, H, N, and U, respectively, designate the Best, Mean Quality, Highest Quality, Number, and Unusual scores.

Table 5
Varimax Factor Loadings and Extension Variable Loadings Deleting the Unusual-High Score for Each Test[a]

Test	Score	Factor		
		I	II	III
FH	Best	.04	.58	.09
	Mean Quality	-.14	.57	.03
	Highest	.14	.64	-.09
	Number	.80	.40	-.04
	Unusual	.68	.17	-.05
EP	Mean Quality	-.19	.53	.55
	Highest	.15	.35	.73
	Number	.70	.04	.32
	Unusual	.53	-.18	-.09
SMP	Best	.25	.45	.45
	Mean Quality	-.22	.12	.72
	Highest	.33	.14	.71
	Number	.84	-.10	-.18
	Unusual	.40	-.21	-.56
MC	Best	-.01	.64	.13
	Mean Quality	-.10	.64	.18
	Highest	.15	.65	.35
	Number	.61	.06	.24
	Unusual	.55	-.16	-.04
Extension Variables	GRE-V	.27	.39	.41
	GRE-Q	.22	.44	.28
	Adv. Psychology	.23	.30	.47
	Experimental Subscore	.22	.27	.38
	Social Subscore	.18	.26	.47

[a]Loadings of .40 and greater have been underlined.

SMP—had a large negative loading on Factor III.

At least two related possible interpretations of the two quality factors can be offered. The tests which form Factor II may, in effect, be more highly divergent in nature; the problems may impose fewer constraints and thus allow the examinee to draw from a wider range of possibilities than those associated with Factor III. Alternatively, the tests comprising Factor III may place more emphasis on issues of the design and analysis of experiments than do those associated with Factor II.

Oblique rotation of the three-factor solution produced highly similar results. The two quality factors obtained by this method correlated .27, indicating only a slight degree of relation between them. At the same time, each was essentially independent of the number factor.

Analyses were also carried out using the full 23 × 23 matrix. Addition of the four Unusual-High scores did not alter the configuration in any substantial way. These scores did not load consistently or heavily on any of the factors.

When more than three factors were retained, the results became more nearly test-specific. With five factors, for example, three quality factors were obtained, matching the results given by the Smallest Space Analysis. There were also two factors loaded only by the count scores; but no pair of tests was either separated or linked consistently in the structure.

At the bottom of Table 5 are given loadings for the GRE scores on the three factors derived from that analysis; these loadings will be discussed in the following section.

The results of these factor analyses were quite consistent with those obtained by Smallest Space Analysis. First, the several quality scores from a given test, even when statistically freed of experimental interdependencies, consistently appeared together on the same factor. Second, response quality did not represent a single underlying ability dimension across the set of tests. There were at least two, and possibly three, distinct abilities. Quality of ideas on EP and SMP defined a single factor throughout the various analyses, contrasting with quality of ideas on FH and MC. Third, in the count scores for each test, the Number and Unusual scores tended to cohere. The results were similar to those obtained from studying performance on simple ideational fluency tasks (e.g., Ward, 1969), in which the number of unusual responses appears to be a direct consequence of the rate at which the more obvious possibilities are exhausted, rather than representing a distinct process in itself. Finally, the bases on which individuals were differentiated in idea quality were unrelated to the bases on which they were distinguished in quantity of solutions.

Correlations with Other Measures

Correlations with GRE Scores

Correlations with GRE aptitude and Advanced Psychology scores are of significance in relation to the construct and discriminant validity of the experimental tests. The new tests presumably require verbal comprehension and expression of ideas, reasoning from the terms of the problem given, and some understanding of facts and principles in psychology. Thus, they may be predicted to bear some relation to GRE Verbal, Quantitative, and Advanced Psychology scores. Moreover, differences in the magnitudes of their relations to these scores may provide an indication of the abilities or processes involved in solving these problems in scientific thinking. On the other hand, the new tests must be discriminable from all the more conventional assessment indices in order to be useful, either practically or in research.

Estimated correlations with GRE scores are presented in Table 6. These coefficients were derived from the reliability sample only; with this group test variances could be estimated without resorting to the correction procedures required for the intercorrelation sample. For correlations involving scores from the Advanced Psychology Test, *N*'s approximate those reported in Table 1. For the aptitude scores, *N*'s are reduced by about 19%, since a number of individuals did not take the Aptitude Test in the October 1973 administration.

In general, quality scores from the experimental tests were more closely related to the GRE measures than were the count scores. This finding is consistent with studies using an earlier version of FH (Frederiksen & Evans, 1974) and with the results of the factor analysis. However, examination of estimated true-score correlations (not shown) indicated that a substantial amount of the true variance in the experimental scores was not accounted for by any GRE test. For example, the median true-score correlation of Verbal Aptitude with quality scores over all the tests was .44, equivalent to an overlap of 19% in variance. For the count scores, the comparable median coefficient was .20, corresponding to an overlap of only 4% in true variance.

Within rows in the table, correlations for the Verbal score were typically highest, while correlations for the two subscores of the Advanced Psychology Test were usually lowest. Reasoning, as represented by the Quantitative aptitude test, had slightly greater relations with the new scores than those found for the Advanced Psychology Test total score. However, the median coefficients for the five columns ranged only from .24 to .28. This striking similarity in the correlations might indicate that the new tests depend on verbal ability, reasoning, and knowledge of psychology in equal degrees. These coefficients might also reflect the fact that the GRE scores do not represent pure cognitive abilities.

SMP appears to be the test which most overlaps with GRE tests in variance, while MC appears to be the test with the least overlap in variance. SMP is also the least reliable of the experimental tests: its scores have the most erratic relations with those of the other tests. In its present form, therefore, it may be less useful than the other new instruments. However, it is possible either that new items could be created which

Table 6
Correlations with GRE Scores[a]

Score	Test	Verbal	Quanti-tative	Advanced Psychology	Experimental Subscore	Social Subscore
Best	FH	.34	.34	.28	.24	.25
	EP	--	--	--	--	--
	SMP	.39	.32	.40	.33	.42
	MC	.28	.30	.24	.24	.18
Mean	FH	.31	.28	.24	.20	.24
	EP	.37	.31	.33	.29	.31
	SMP	.39	.34	.45	.36	.46
	MC	.29	.34	.25	.24	.18
Highest	FH	.38	.37	.31	.26	.31
	EP	.47	.40	.42	.37	.39
	SMP	.49	.39	.53	.50	.46
	MC	.35	.33	.31	.25	.29
Number	FH	.25	.35	.22	.21	.17
	EP	.35	.27	.29	.27	.22
	SMP	.15	.14	.13	.19	.02
	MC	.22	.05	.19	.10	.27
Unusual	FH	.13	.20	.12	.13	.09
	EP	.13	.09	.07	.09	.01
	SMP	.01	.03	-.09	-.03	-.14
	MC	.28	.10	.23	.18	.26
Unusual-High	FH	.19	.19	.17	.14	.16
	EP	.15	.13	.11	.15	.03
	SMP	.20	.21	.18	.21	.10
	MC	.24	.21	.32	.26	.32

[a]Based on the reliability sample.

would better meet the requirements of construct and discriminant validity or that SMP might be more appropriate for students more advanced in training and experience.

GRE scores were carried as extension variables in the factor analysis (i.e., they were not allowed to influence the factor structure); the loadings are shown at the bottom of Table 5. All the GRE scores had similar small relationships with the factor representing numbers of responses. Correlations with Factor II were somewhat higher, especially for the two aptitude tests; correlations with Factor III were even higher, especially for the Advanced Psychology Test. Although the differences in loadings were too small to permit any concrete generalizations, it would appear that EP and SMP require more knowledge of psychological facts and principles than do FH and MC. In turn, FH and MC require more reasoning.

Re-analysis for Select Sample

A select sample was chosen, consisting of 32% of the reliability sample. It included those individuals who, according to the questionnaire filled out at the time of testing (1) were seniors; (2) were majoring in psychology; (3) had training in statistics; and (4) earned scores of 510 or higher on the Verbal aptitude test. The major parts of the data analysis were repeated for the members of this group.

In almost all instances, the means on the experimental tests were higher for the select sample than for the complete reliability sample; and the standard deviations were slightly smaller. Correlations of scores with those on the GRE tests were appreciably lower for this group; for example, the median of the correlations with GRE-Verbal was .13, as compared to .28 for the reliability sample. These differences are consistent with expectations for a group which is restricted to the upper portion of the range of talent in the sample.

Comparison of lower-bound reliability estimates showed few consistent differences between the two groups. Although for FH quality scores reliabilities were lower in the select group, SMP reliabilities tended to be slightly higher. It is possible that SMP was not measuring the same abilities in the two groups because the test was too difficult for the poorest candidates. In spite of some restriction in range, reliability was in general approximately as good for the select sample as it was for the larger sample.

The select group was also compared with the complete reliability sample with respect to the standard errors of measurement. In 20 of the 23 comparisons, the error of measurement was smaller for the select sample. The largest differences were for quality scores. For FH Highest Quality, for example, the standard errors of measurement for the reliability and the select samples were 1.72 and 1.36, respectively. Thus, the accuracy of measurement was actually higher for the select sample.

The new tests did not duplicate the existing GRE tests with regard to abilities measured, and there was a substantial amount of true variance not predictable by the GRE battery. The variance in experimental tests predictable from GRE tests for the select sample was appreciably smaller than for the complete sample, although, in general, reliabilities were relatively unaffected by the restriction in range of ability.

Predictive Validity

In April 1975, when the typical examinee should have been finishing his/her first year of graduate study, students were contacted by mail and asked to complete a questionnaire dealing with their academic activities, interests, and accomplishments during the current school year. The purpose of this follow-up was to investigate the ability of both the Tests of Scientific Thinking (TST) and GRE scores to predict various indices which might serve as both traditional and non-traditional criteria of successful performance at an early stage in a scientific career.

Altogether, 42 variables were generated from the questionnaire items. These were grouped into six general areas: (1) indices of student and department quality; (2) program emphasis; (3) areas of professional interest; (4) single preferred professional activity; (5) self-appraisals of skills and knowledge; and (6) professional activities and accomplishments during the first year of graduate work.

Questionnaires were mailed to the 3,235 individuals who had given permission to be contacted for a future investigation. A larger number of cases than expected were lost, either because questionnaries were not deliverable or not returned or because students were not enrolled in a graduate program in psychology. Of the questionnaires presumably delivered, about 50% were returned; approximately half of those returned reported attendance in a psychology program. After this attrition, 654 cases remained. However, only 251 of these were students who had been part of the reliability sample in the original study, and it was this sample which had proved to be the most useful group for the study of predictive validity. The various sources of reduction, no doubt, produced data for which the assumptions of the item-sampling method no longer held, as well as reducing the number taking any one test to about 60.

Based on their responses to the questionnaire, the students in the follow-up sample can be characterized as follows: (1) almost 90% aimed for a Ph.D. rather than a master's degree; (2) the students earned high grades both as undergraduates and as graduate students; (3) many were interested in clinical practice, applied research, and teaching, while few were interested in administration or guidance and counseling (when asked to choose one area, 42% said they preferred clinical practice); (4) they rated themselves high in knowledge of psychology and statistics and low in teaching and clinical ability; (5) most had a fellowship or assistantship; (6) finally, many of them had experience in research during the first year of graduate work, while few were involved in writing or editing.

Several comparisons were made between the follow-up sample and the original GRE sample. The two groups did not differ appreciably with regard to scores on the TST. As expected, upon examination of the GRE scores, the follow-up sample was found to be select. For the V, Q, and Advanced Psychology test the mean scores were 585, 585, and 591, respectively, for the students who reported they were attending a graduate psychology program; the means were slightly more than a third of a standard deviation above those for the entire original GRE group (558, 549, and 552, respectively). However, the variability of test scores showed very small differences between the two groups. For the original sample the standard deviations of the three GRE test scores were 100, 118, and 93; for the follow-up group the corresponding values were 95, 108, and 88. Thus, despite the selectivity operating, restriction of range should have only a minor effect on correlations with the questionnaire variables.

Lower bound reliabilities of scores from the TST were calculated for the 251 follow-up cases which had been part of the reliability sample. The coefficients were, in general, higher than those originally obtained; however, they varied somewhat erratically. Reliabilities were recalculated both for the follow-up and the original GRE sample, using coefficient alpha. Alpha was calculated using inter-item covariances, rather than their squares, which are required for Lambda 2. It should, therefore, have been less sensitive to distortion resulting from variability introduced by small sample sizes and possible violations of item-sampling assumptions. On the average, these estimates were quite similar for the two samples; their mean difference across all scores was only .02.

Finally, correlations of GRE with experimental test scores were recalculated. The correlations were, in general, similar for the original GRE sample and the follow-up sample, although they were slightly lower (.03 on the average) for the latter group. There was some tendency for the Number scores to have higher correlations and the Quality scores to have lower correlations for the follow-up group. The exception was SMP, which had higher Quality score correlations.

In general, while the follow-up sample was select in terms of GRE scores, its performance on the TST was quite similar to that of the larger group from which it was drawn. Apparently, abilities reflected by these tests did not enter appreciably into graduate school selection decisions.

Table 7 shows the correlations between GRE scores and the 42 questionnaire variables. Since the *N*'s are large, sample size and item sampling are not of concern. A large proportion (55%) of the coefficients in Table 7 were significant at the .01 level; however, with the large *N*'s a statistically significant correlation may be no higher than .10. Correlations of .25 or higher are underlined to call attention to the relationships that are most likely to be useful.

Only 17 correlations between GRE scores and questionnaire variables equalled or exceeded .25. These were concentrated, with two exceptions, in the portion of the table representing indices of student and department quality. Furthermore, most of these (10 of the 15) were with the two indices of department quality—whether or not the department has APA accreditation (APA, 1974) and a rating of department quality (Roose & Andersen, 1970). Other correlations underlined indicate that high GRE scores tended to be associated with having a fellowship or other support; but GRE scores were very likely involved in making these awards.

Only two correlations equalled or exceeded .25 in the remainder of the table. The first was a correlation of −.29 between the experimental psychology subtest and emphasis on the practice of psychology in the training program. The second was a correlation of .25 between the same subtest and self-appraisals of ability to interpret research results. In interpreting this correlation, it must be taken into consideration that GRE scores were known by the students. None of the GRE tests correlated more than .18 with reports of professional activities.

If all the statistically significant correlations are considered, it appears that the relationships generally (but not always) appear to make theoretical sense. For example, all except one of the correlations with "Plans academic program" were positive, and all correlations with "Program objective: practice" were negative. All the correlations with interest in applied research and basic research were positive, while all those with interest in guidance and counseling were negative. All the correlations with self-appraisal

Table 7

Correlations of Questionnaire Variables with GRE Scores[a]

Questionnaire Variable	V	Q	Adv.	Exp.	Soc.-Pers.
Student/Dept. Quality					
2. Plans PhD rather than MA	.19	.12	.18	.17	.17
3. Undergraduate GPA	.22	.14	.20	.18	.19
4. Graduate GPA	.18	.15	.26	.22	.23
5. Attends accredited dept.	.38	.27	.34	.32	.29
29. Rates dept. high in quality	.27	.23	.18	.18	.14
30. Satisfied with department	.17	.14	.10	.09	.10
31. Department quality index	.37	.31	.30	.28	.25
32. Support through psychology	.22	.17	.27	.29	.19
33. Support through fellowship	.27	.15	.17	.13	.19
Program Emphasis					
6. Plans academic program	.14	.15	.11	.20	-.01
27. Program objective: practice	-.19	-.17	-.23	-.29	-.12
28. Deemphasizes teaching	.03	.02	.04	.02	.04
Interest Areas					
7. Number of areas	.12	.01	.13	.12	.11
8. Administration	-.06	.00	-.01	-.01	-.02
9. Applied research	.19	.12	.16	.18	.10
10. Basic research	.14	.12	.17	.23	.06
11. Clinical practice	.01	-.06	.00	-.08	.13
12. Guidance and counseling	-.15	-.21	-.22	-.23	-.14
13. Teaching	.09	.06	.16	.16	.13
Preferred Area					
14. Administration	-.08	-.02	-.07	-.05	-.09
15. Applied research	.06	.02	.02	.06	-.04
16. Basic research	.10	.10	.14	.19	.07
17. Clinical practice	-.03	-.03	-.03	-.11	.08
18. Guidance and counseling	-.13	-.13	-.17	-.17	-.13
19. Teaching	.03	.05	.05	.06	.02
Self-Appraisal					
20. Mean rating	.04	.03	.17	.18	.13
21. Knowledge of psychology	-.02	-.08	.14	.17	.07
22. Knowledge of statistics	.07	.21	.15	.16	.07
23. Experimental design	.05	.01	.11	.13	.08
24. Research interpretation	.15	.12	.23	.25	.15
25. Clinical ability	-.10	-.17	-.12	-.20	.02
26. Teaching ability	-.02	-.03	.08	.07	.07
Professional Activities					
34. Number of activities	.01	.02	.13	.19	.04
35. Meetings, subscriptions	-.12	-.17	-.02	-.03	-.01
36. Publications	.04	.00	.16	.18	.12
37. Planned, did indep. research	-.03	-.05	.01	.04	-.01
38. Did collaborative research	.06	.04	.13	.17	.05
39. Taught undergraduates	.07	.03	.11	.13	.07
40. Advised on statistics	.01	.09	.02	.04	-.03
41. Helped prepare book	.12	.05	.05	.05	.03
42. Worked with equipment	.00	.13	.09	.18	-.04
Other					
1. Has MA rather than BA	-.07	-.19	.11	.06	.14

[a] For correlations with V and Q scores, *N*'s are usually about 525, and for the Advanced Psychology Test scores they are usually about 650. With *N* of 525, a correlation of .09 is significant at the 5% level, and one of .11 is significant at the 1% level. Coefficients of .25 or higher are underlined.

of research interpretation ability were positive, while all but one with appraisal of clinical ability were negative. However, no consistent relationships were found for the "professional activities" variables.

It was originally intended that the correlations between TST scores and questionnaire variables would be computed using the estimation procedures for the item-sampled data previously described. However, as has been seen, reliability coefficients for the follow-up sample were somewhat unstable, presumably because of small sample sizes coupled with the likelihood that the assumption of random assignment to item-sampled subgroups does not hold. It is therefore probable that the estimated correlations with other variables would also be unstable; furthermore, no test of significance would be available. It was, thus, necessary to find a method not requiring the use of item variances and covariances to show relationships of experimental-test scores with questionnaire variables. The reliability sample provided the data for an appropriate method.

Each student in the reliability sample took a subtest composed of three items from one test. Since there were 20 possible three-item combinations of six items, only three students, on the average, took exactly the same test. If it is assumed that the items are interchangeable (after adjustment for differences in item means), scores can be obtained which are based on whatever three-item test was taken. The correlations based on these scores may be more stable than those derived from the estimation procedure, and they can be tested for significance by conventional methods. The correlational results which follow are based on such three-item tests.

Each item score was adjusted by subtracting the grand mean obtained from the complete reliability sample in the original study. Since sample sizes and item-sampling procedures in the original study gave reasonable assurance of equivalence of ability for subgroups, this procedure should have provided an appropriate adjustment for differences in item difficulties. After this adjustment, six test scores were obtained for each student by summing over the three-item test he/she had taken. These scores were then correlated with the 42 questionnaire variables. A correlation of .25 is significant at the .05 level of confidence.

A count of the number of coefficients attaining the .05 level for different kinds of scores is revealing. There were a total of 462 correlations between questionnaire variables and *quality* scores, of which 21 were significant. Since the expected number of correlations significant by chance is 23, it cannot be concluded that any real relationships exist. For the remaining three scores—Number, Number of Unusual, and Number of Unusual-High Quality Responses—the situation was slightly better: Of 504 coefficients 37, or 7.3%, reached significance at the .05 level. There is evidently some non-chance covariation in this part of the data, but it is heavily embedded in noise.

In the previous discussion of reliability and corrections, six-item tests were assumed in order to have more adequate reliability while remaining within a reasonable time requirement. In contrast to the three-item tests reported on immediately above, more correlations reaching significance would, of course, be expected with a longer test. The procedure previously described for estimating correlations for a six-item test was applied to the data. The number of correlations reaching significance was then determined. For the quality scores 29 (6.3%) were found, and for the count scores 75 (15.1%) were significant. Thus, there was evidence of real relationships between the questionnaire variables and the three scores based on number of responses.

A method was obviously needed to reduce the number of coefficients to be examined and to focus attention on the relationships that are most meaningful. The use of r to z transformations for averaging is appropriate if the assumption is made that the four test situations—FH, EP, SMP, and MC—constitute four replications of the same experiment. The finding that the quantity scores for all four tests loaded on one factor may be taken as evidence supporting this view. A simple significance test for the average r is available (McNemar, 1962).

Table 8 provides a summary of the significant coefficients obtained by this procedure. Number, Number of Unusual, and Number of Unusual-High Quality Responses showed, respectively, 7, 9, and 10 relationships significant at not more than the .05 level, where for each score 2 would be expected by chance.

There is evident consistency across the three scores in the location of significant results within

Table 8
Significant Average Correlations Combining
Data over Four Tests[a]

Questionnaire Variable	Score		
	Number	Unusual	Unusual-High
Student/Dept. Quality			
2. Plans PhD rather than MA	.18**	.18**	.16*
31. Department quality index	.20**		.17**
32. Support through psychology	.17**	.15*	.15*
Program Emphasis			
6. Plans academic program			.14*
27. Program objective: practice			-.20**
Self-Appraisal			
20. Mean rating		.13*	
23. Experimental design	.15*	.17**	
25. Clinical ability			-.13*
Professional Activities			
34. Number of activities	.18**	.24***	.16*
35. Meetings, subscriptions		.14*	
36. Publications		.18**	.13*
38. Did collaborative research	.19**	.18**	.15*
40. Advised on statistics	.13*		
42. Worked with equipment		.19**	.19**

[a] Two-tailed probability levels.
*p < .05
**p < .01
***p < .001

sections of the table. Of the 26 average correlations significant at not more than the .05 level, 12 were derived from questions grouped under "Professional Activities." Students scoring high on the experimental tests tended to have engaged in more of these activities. Specifically, they tended to have attended professional meetings and subscribed to journals, published, engaged in collaborative research, provided advice on experimental design and statistics to other students, and worked with laboratory equipment.

Eight significant relationships were found within questions grouped under "Student Department Quality." Students scoring high tended to have plans for the Ph.D. rather than a master's degree, to attend a department with a high Roose-Andersen quality index, and to have obtained their support through a psychology-related activity during the first year. Four more coefficients were found within the section of the table designated "Self-Appraisal." Students scoring high on number of Unusual responses generally rated themselves high. Along with those students scoring high on Number of Responses, they claimed to have knowledge of experimental design. Those scoring high on Number of Unusual-High Quality ideas viewed themselves as having a lower level of clinical skills. Finally, there were significant relationships of Number of Unusual-High Quality ideas to two indices indicating enrollment in a program emphasizing research rather than practice.

A similar analysis was carried out on correlations of questionnaire variables with the three quality scores. In view of the chance number of significant relationships seen for individual correlation coefficients, it was not expected that positive results would be found. And, in fact, only four relationships reaching the .05 level were found; the chance expectation would have been six.

In spite of the low reliability of the TST scores, these results indicate some ability for certain of the scores to predict possibly important aspects of first-year graduate performance. Additionally, they suggest a discrimination between the experimental tests and the conventional predictors regarding the aspects of the performance to be predicted. In examining correlations between the Advanced Psychology test and the questionnaire variables, substantial relations were found primarily with variables representing individual or departmental quality indices; there was little relation with variables representing professional activities.

Summary and Discussion

The purposes of this investigation were (1) to develop a set of tests that might reasonably be used as criterion measures in research on scientific thinking—particularly creative thinking—and (2) to assess the suitability of these tests for use as criterion variables from the standpoint of their psychometric properties. The needed evidence was composed of test difficulty, reliability, intercorrelations, and validity.

The tests seemed to be sufficiently challenging, in terms of difficulty, for graduate students, even for a select subsample of applicants for admission to graduate school.

Scores derived from each of the tests formed two major clusters: one reflecting quality and the other quantity of ideas. For the quality scores and for Number of Responses, reliabilities for a six-item test were quite adequate for a research instrument. The two remaining scores, Number of Unusual Responses (which fell into the quantity cluster) and Number of Unusual-High Quality Responses (which did not have consistent relations to other scores), had lower reliabilities, especially for two of the tests. Since a six-item test can be given in less than 50 minutes, it would be feasible to employ still longer tests and, consequently, to increase the reliabilities of all the scores.

The tests do nct merely reflect abilities already measured by the GRE aptitude and achievement tests. The scores based on Number of ideas and, in particular, Number of Unusual ideas were unrelated to the conventional tests. The quality scores had some correlation with GRE measures, as would be expected on theoretical grounds. However, even if all the tests were to be perfectly reliable, only a small proportion of the variability in quality scores would be predictable from GRE scores. Thus, the Tests of Scientific Thinking do have reliable variance that is not being measured by the conventional tests.

In view of the low rate of questionnaire returns in the follow-up study, the deficiencies of questionnaire responses as criterion variables, and the incomplete data resulting from the use of item sampling, evidence of predictive validity is perhaps more encouraging than might be expected. The most valid scores were those based on counts—Number, Number of Unusual, and Number of Unusual-High Quality Responses. These scores were significantly related to some questionnaire variables in the area of student and department quality—an area where the GRE tests had their highest validity. In addition, there is evidence that these scores were related to self-appraisals of professional skills and to reports of professional activities and accomplishments—areas where the GRE tests have low validity. These questionnaire variables would appear to be reasonably good early indicators of productive scientific work, even though they were obtained at the end of the first year in graduate training.

Further investigation seems warranted, using complete tests and criterion variables based on professional careers of subjects. Other studies should be directed primarily at questions of construct validity. Is performance on the tests related to other personal characteristics in ways that are theoretically consistent? Does performance change in relation to experimental or educational grounds? Answers to such questions will not only provide evidence on validity, but also contribute to an understanding of processes involved in scientific thinking.

References

American Psychological Association. APA-approved doctoral programs in clinical, counseling, and school psychology: 1974. *American Psychologist*, 1974, *29*, 844–845.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, *16*, 297–334.

Flanagan, J. C. et al. *Critical requirements for research personnel: A study of observed behavior of personnel in research laboratories.* Pittsburgh: American Institute for Research, 1949.

Frederiksen, N., & Evans, F. R. Effects of models of creative performance on ability to formulate hypotheses. *Journal of Educational Psychology*, 1974, *66*, 67–82.

Guilford, J. P. Some new looks at the nature of creative processes. In N. Frederiksen & H. Gulliksen (Eds.), *Contributions to mathematical psychology.* New York: Holt, Rinehart & Winston, 1964.

Guilford, J. P. *The nature of human intelligence.* New York: McGraw-Hill, 1967.

Guion, R. M. Content validity—the source of my discontent. *Applied Psychological Measurement*, 1977, *1*, 1–10.

Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

Guttman, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, *10*, 255–282.

Guttman, L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, 1968, *33*, 469–506.

Koutsopoulos, C. J. *The mathematical foundations of classical test theory: An axiomatic approach II* (Research Memorandum 64–3). Princeton, NJ: Educational Testing Service, 1964.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 1968.

MacKinnon, D. W. The nature and nurture of creative talent. *American Psychologist*, 1962, *17*, 484–495.

McNemar, Q. *Psychological statistics* (3rd ed.). New York: Wiley, 1962.

Roose, K. D., & Andersen, C. J. *A rating of graduate programs.* Washington, DC: American Council on Education, 1970.

Ward, W. C. Rate and uniqueness in children's creative responding. *Child Development*, 1969, *40*, 869–878.

Acknowledgments

We are grateful to the Graduate Record Examinations Board for their support of this research and to Robert Altman and Martin Glaubitz for their help in arranging for the test administration. Two research assistants, Suzanne Taweel and Charlotte Kiefer, served with us as judges, aided in the development of the scoring procedure, supervised the scoring, and contributed to the research in many other ways. Frederic Lord provided useful advice in the development of computational procedures, and Ledyard Tucker suggested several alternative procedures for data analysis.

Author's Address

William C. Ward, Division of Psychological Studies, Educational Testing Service, Princeton, NJ 08540.