Predicting Doctorate Attainment With GRE and Other Variables (1965)

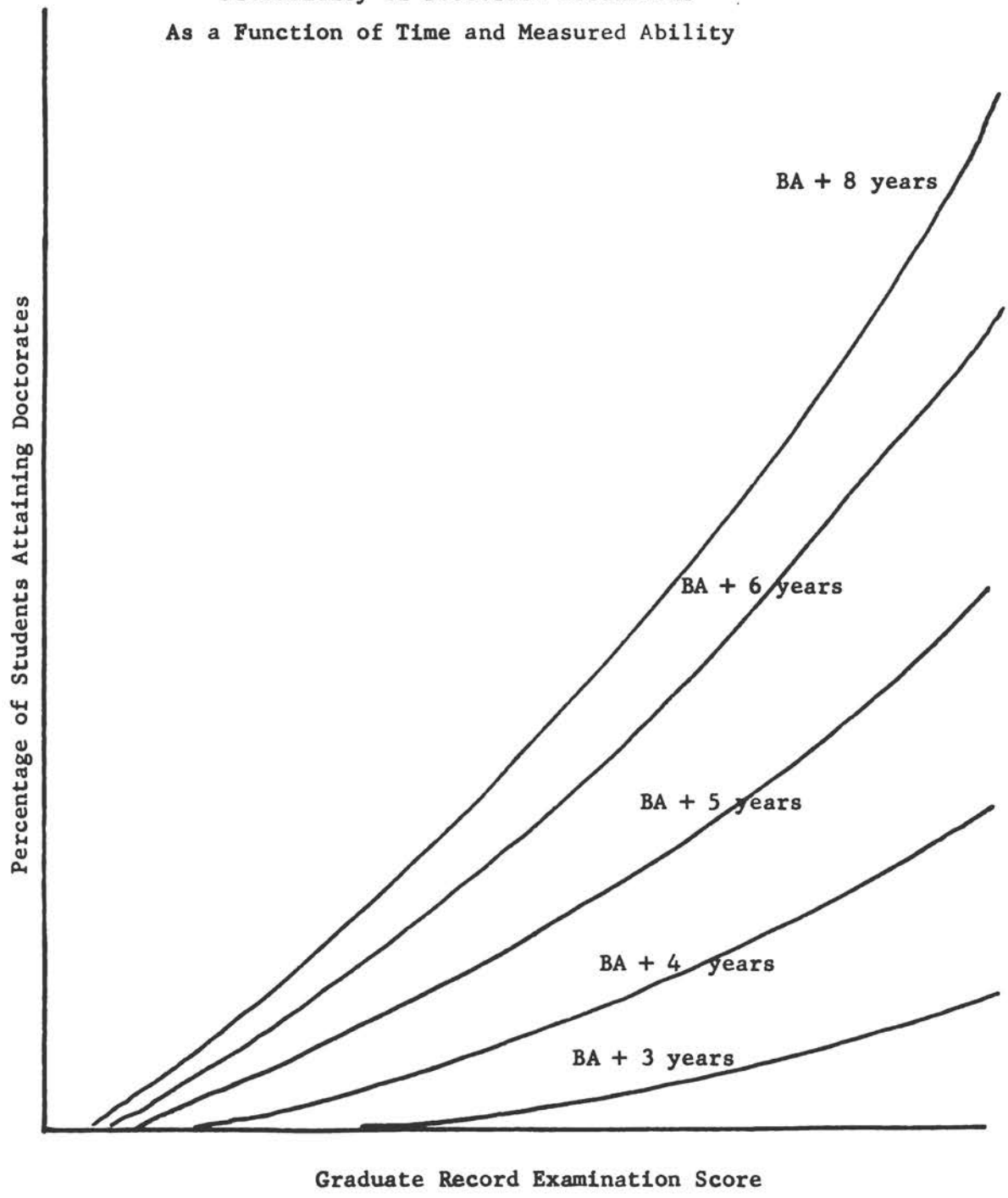
Pages
53

Size
8.5 x 10

ISBN
0309360730

Creager, John A.; Office of Scientific Personnel; National Research Council

Probability of Doctorate Attainment
As a Function of Time and Measured Ability



ABSTRACT

This report describes two closely related studies of the validity of the Graduate Record Examinations and other variables for the prediction of various doctorate attainment criteria. In the first study, using primarily first-year applicants for National Science Foundation graduate fellowships and involving only the Graduate Record Examinations, validation is performed on broad-based samples stratified on Advanced Test norms developed in the National Program for Graduate School Selection. The second study is a part of a follow-up of former applicants for National Science Foundation graduate fellowships, and uses samples of 1955 and 1956 first-year and intermediate applicants. Validation is performed on the fellowship selection battery, which includes the undergraduate science-mathematics grade-point average, ratings from reference reports, and Graduate Record Examinations scores. The study includes validation of various composites of the fellowship selection information.

The studies yield considerable information concerning the significantly useful but limited validity for these variables in predicting doctorate attainment criteria, even within the upper ranges of ability. The predictive power of individual variables and the increase in prediction obtained by combining the variables is shown. In addition, several methodological issues are discussed in connection with various ways of defining and refining doctorate attainment criteria. The results are summarized with recommendations for future studies.

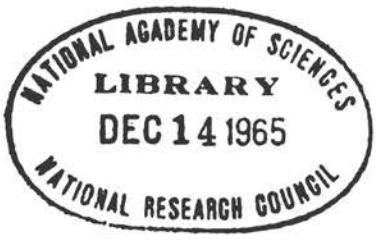


TABLE OF CONTENTS

GRADUATE RECORD EXAMINATIONS AND OTHER VARIABLES IN PREDICTION OF SCIENCE DOCTORATE ATTAINMENT

I. THE BACKGROUND, CONTEXT, AND SCOPE OF THIS REPORT

The nation values a high level of scientific and technical achievement. This value is reflected in the commitment to study the socio-economic, political, and cultural factors involved in maintaining a high output of doctorates in the science fields. Considerable effort and costs are dedicated to the selection and education of those persons with ability to contribute to the nation's scientific and technical achievement and to further education.

Concern and responsibility for successful completion of higher education and for the production of science doctorates is shared by many persons and agencies in the academic community, in science-oriented industry, and in the government. Since 1952 the National Science Foundation has maintained a graduate fellowship program; persons selected for their ability to pursue graduate study in the sciences and later contribute to the national scientific and technical achievement are awarded fellowship support for their graduate study. In conjunction with this fellowship program, the National Science Foundation has also supported a research program designed to improve the procedures used in evaluating fellowship applicants. The two studies described in this report are a part of that research program, but their import bears on the larger context of selection for graduate study in the sciences by other agencies and institutions.

... GRE Validity

In the early years of the fellowship selection research program, emphasis was placed on the description of applicant groups and on studies of the reliability of the evaluation procedures. With the passage of time, many of these applicant groups have completed their formal education and taken up roles in the scientific and technical community. Therefore, in recent years the fellowship selection research program has been oriented toward follow-up of former applicants, to the development and evaluation of various validation criteria, and to the use of these criteria to validate the selection procedures. This report describes two closely related studies carried out in the context and spirit of the follow-up aspects of fellowship selection research. Both studies are concerned with the prediction of doctorate attainment and of the time required to complete a doctorate in those science fields supported by the National Science Foundation. However, the two studies differ in purpose and focus. The first study is concerned with

the validation of the Graduate Record Examinations against doctorate attainment criteria. It aims at a level of generality not heretofore attempted, since most validation studies of these tests have been done on limited samples within individual institutions and for local purposes. In contrast, this study aims at general validation for graduate school selection and admission to higher degree candidacy. It also aims to overcome restriction to particular fellowship applicant groups. Therefore, although the first study uses former fellowship applicants, it is concerned with specially constructed samples, stratified on the national test performance norms. This provides a broader validation base and fuller range of talent than is possible with highly selected fellowship applicant groups.

... Validity of Other Variables

The second study returns to the emphasis on follow-up of former fellowship applicant groups with its focus on the evaluation of the fellowship selection procedures. This study therefore uses subjects involved in an on-the-job follow-up study currently being carried out on the 1955 and 1956 first-year and intermediate applicants for graduate fellowships. In addition to the Graduate Record Examinations, variables validated include the undergraduate science-mathematics grade-point average and ability ratings from the reference reports submitted in support of a fellowship application. A weighted sum of these variables, called a "summary score", and the panel evaluations of the applicants (the quality group ratings) are also included in the validation against doctorate attainment criteria. This study is primarily useful in the fellowship program, in contrast to the more general focus of the first study.

Doctorate Attainment and BA-PhD Time Lapse as Criteria: Their Nature and Import

The decision to admit a student to graduate school, to provide financial support of his graduate education, or to admit him to candidacy for a higher degree implies an expectation that formal graduate education will be completed. The attainment of the doctorate degree is the primary indication that such an expectation has been fulfilled. In the science fields, doctorate attainment is also a prerequisite for most positions of influence on further development of science.

... BA-PhD Time Lapse as a Criterion

Among those who do attain a science doctorate, the time required to do so is of considerable interest. The Office of Scientific Personnel of the National Academy of Sciences, through its Doctorate Survey, has obtained information on the time lapse between acquisition of the baccalaureate and doctorate degrees. Informal checks for subjects included in these studies show that time out of study between attainment of the baccalaureate degree and entry into graduate school is rare. It is probably more common in the general graduate student population than among fellowship applicants. Interruption of graduate education for an appreciable period is also not very common in the subjects studied here. Appreciable numbers do report part-time graduate study, presumably while working on teaching assistantships, or in external employment. To the extent that assistantship status may be related to ability, controlling for it in validation may not be too critical. To the extent that other interruptions are not related to ability and are culturally determined factors, over which there is only limited possible control, concern about them in a validation study is irrelevant. Therefore, only the gross BA-PhD time lapse in years is used in the studies described in this report.

... How Many Make It?

Considerably greater concern about time lapse information in validation studies is involved in defining a doctorate attainment criterion. It is rare that one completes a doctorate within three years after the baccalaureate. In the science fields the mean time lapse is approximately eight years, with greater deviations above the mean than below. If time were allowed for everyone to complete a doctorate who would in fact eventually do so, one would have to wait too long to do the study. Not only would more of the people attaining doctorates have more time out and extensive part-time study (thus complicating the interpretation of results), but more persons of low measured ability would have achieved a doctorate under possibly lower standards of dissertation and course quality. Too short a time lapse would eliminate many high quality people, especially those very able persons who take on more difficult dissertations and a more enriched course program. These considerations lead, for criterion definition purposes, to setting limited cutoff times for doctorate completion, as described below.

... August 1964 as cutoff

Three indicators of doctorate attainment were used in the first study and two of these in the second study. The first indicator is based on completion of a doctorate in the United States up through August 1964. It is assumed that foreign doctorates were too rare to have introduced any serious bias. Since most of the subjects used applied for first-year or intermediate fellowships in 1955 and 1956, they have had eight or nine years from fellowship application time and not more than ten years from baccalaureate completion to attain a doctorate. For the science fields involved in these studies, this is slightly longer than the mean time lapse for these fields. From this definition of doctorate completion, the first doctorate attainment criterion was derived as a completion-noncompletion dichotomy.

... Time Lapse Coded

The second indicator of doctorate attainment was obtained by coding the time lapse as follows, using the same definition of completion:

BA-PhD Time Lapse (in years):	Less than 4	4	5	6	7	8	9	No PhD by Aug. 1964
Coded Variable:	1	2	3	4	5	6	7	8

Where the coded variable is used as a criterion, it is considered as a continuous variable grouped in ordered categories for validation purposes.

... Above/Below Field Median

The third indicator of doctorate attainment, used only in the general GRE validation study, dichotomizes the time lapse at the next lower integral value below the mean time lapse for each field of doctorate. This procedure, which comes close to defining doctorate attainment at the field medians (instead of the means) tends to remove field differences in time lapse average and skewness, thus making comparisons of validities across fields somewhat more trustworthy. Comparison of the validities against this third indicator with the validities against the first indicator permits examination of the effect of criterion definition within each field.

... Previous Findings Summarized

More detailed information on time lapse and doctorate completion rates may be obtained from the NAS-NRC Publication 1142 (1) and from Technical Reports #18 and #24 in this series(2,3). As a part of the background for the two studies described in this report, the following statements summarize the relevant findings and their implications for the present studies:

a. First-year fellowship applicants usually apply for a fellowship during their senior year and enter graduate study the same year as that in which they receive their baccalaureate degrees. Intermediate applicants are those who will have completed one full year of graduate study before the beginning of the fellowship year. Thus, for a given application year and fixed cutoff time, the BA-PhD time lapse will be similar for first-year and intermediate applicants, but the time lapse between application time and doctorate will be shorter on the average for the intermediate applicants. In addition, the doctorate attainment rates will usually be higher for the intermediate applicants than for the first-year applicants, due to the selective effects of academic attrition having already occurred to some extent prior to application.

b. Almost all of those applying for fellowships for their terminal year eventually attain doctorates. Thus there is no criterion variance against which predictors could be validated. It is only the first-year or intermediate candidates who experience sufficient attrition to make a doctorate-attainment criterion meaningful. The cases selected for the present study are thus at an optimal level for the validation of selection instruments.

c. Doctorate completion is more likely and requires an average of about one year less time for those who receive NSF fellowship support, than for nonawarded applicants. When ability is controlled, this time lapse difference is reduced nearly one-half. Since there are proportionately more NSF awardees in the samples used in these studies than in the general population of graduate school applicants, doctorate attainment should occur on the average somewhat earlier than in the general population of graduates. In fact, for the subjects used in these studies, peak doctorate attainment occurred in the years 1958 and 1959, with subsequent tapering off of attainment rates in later years. This is further evidence of the desirability of using the August 1964 cutoff date in defining the attainment criterion in these studies.

d. Field differences in doctorate completion rates and time lapse have been rather consistently observed. These differences are due only partly to differences in ability patterns; they are also due to differences in course and dissertation requirements, and to differences in the nature and extent of time-consuming laboratory work, including acquisition and calibration of equipment. Generally the chemists tend to early completion and have high attainment rates. Attainment rate is also high but time lapse somewhat greater for biologists. Time lapse is also greater for psychologists and mathematicians. Completion rates tend to be lower in engineering, mathematics, and the earth sciences.

e. The general picture on sex differences in doctorate attainment is that female subjects are less likely than male students to enter graduate study and less likely to attain science doctorates. They are concentrated in the biological sciences, and, if they complete doctorates, have higher average scores on ability tests.

The Graduate Record Examinations: Uses, Norms, Reliability, and Prior Validation

The Graduate Record Examinations have been widely used for many years. Major uses have been for admission to graduate study, admission to degree candidacy, education program guidance, and fellowship candidate evaluation(4). The examinations include an Aptitude Test, yielding scores in verbal and quantitative abilities; the three Area (cultural background) Tests in social science, humanities, and natural science; and nineteen Advanced (achievement) Tests, each one specific to a given field of study. These tests are administered in two programs, the Institutional Testing program in which the tests are administered to groups within institutions, and the National Program for Graduate School Selection. The National Program provides normative information in terms of which applicants for admission to a particular institution or support program may be appraised, and also provides a basis for the appraisal of students from varied undergraduate backgrounds(5,6).

... 90+% of Fellowship Applicants Have GRE

In the National Science Foundation Graduate Fellowship Program, scores on the Aptitude Test and on the Advanced Test appropriate to the student's field of study are submitted in support of a fellowship application. Over the years

of the fellowship program about 85% of the applicants have submitted one or more GRE scores with their first application. Scores are available for 95% in the years included in the present studies, and for 90-95% in all but the earliest years of the program. The Area Tests are not used; therefore, these studies are confined to the Verbal Test score, the Quantitative Test score, and scores on the Advanced Tests in those science fields supported by the NSF Graduate Fellowship program. The availability of these test scores, having at least face validity for successful performance in graduate school, is most helpful to the fellowship evaluation panels, when the scores are interpreted against the appropriate norms.

... Applicant vs. National Percentiles

In the fellowship program, the scaled scores based on national norms are used, and a percentile based on that year's fellowship applicant group is made available to the evaluating panels. Comparison of such fellowship-applicant based percentiles with national norm percentiles consistently reveals the higher average quality of the fellowship applicant group. Therefore, it is not uncommon for a person with a given scaled score to have moderately high centile rank on the national norms, but a moderately low rank within the fellowship applicant group. Selection of fellowship applicants for awards is necessarily made within the applicant group and, by law, selection is based on "ability". The National Science Foundation bases its decision to award a fellowship primarily on the ability of the applicant relative to his standing with other applicants, as that ability is evaluated by the fellowship panels. This evaluation is <u>not</u> based solely on the GRE test scores. As only a minority of applicants receive awards, it is inevitable that some individuals with one or more GRE test scores above the national average will fail to receive fellowships. Some of these will also attain doctorates and later contribute to the national scientific and technical achievement.

... GRE Reliability

Internal consistency estimates of test reliability have been reported for the Graduate Record Examinations(5) and are well within the usual range of reliabilities for aptitude and achievement tests. They are also well above the minimum required for use in selection. Typical reliability coefficients for the GRE are .90 for the Verbal Test, .84 for the Quantitative Test, and .85-.95 for the Advanced Tests.

... Previous Validity Studies

Since the first of the studies presented in this report is entirely a validity study of the Graduate Record Examinations, and the second study includes a large GRE validation component, it is useful to note to what extent these tests have been previously validated. Considering the widespread use of the GRE, the numbers of persons who have been tested and evaluated on the basis of their performance on these tests, it is surprising that the psychometric literature contains so few validity studies of these tests. It is not uncommon for a particular department or institution to execute one or more small GRE validity studies. The test publisher has summarized some of these validity studies(7,8). In addition, the Air Force has used the tests in appraising the educational achievement of student officers(9).

These studies are typically carried out using the small numbers of cases available within a particular department or institution. Course grades or faculty ratings are used as criterion measures, since these are immediately available as interim criteria on graduate students during the course of their graduate education. This is in contrast to the longer time required for the maturation of a doctorate attainment criterion. Apparently many studies are undertaken to obtain a quick answer to an essentially local question. Although some studies are well designed, many are so markedly restricted in the range of talent studied, through limitations inherent in availability of subjects, as to seriously limit the generalizability of the conclusions. Although these local studies are useful to the departments and institutions that undertake them, the results are not widely reported, nor can be of any but limited use to other departments, institutions, or programs.

... So We Did It This Way

It is such considerations as discussed above that led to the design and execution of the general GRE validation study presented in Part II of this report. In spite of some limitations in using the fellowship applicant samples in this and in the second study, reported in Part III, the subjects come from a fairly wide range of institutions and departments, and cover a wide range of talent. In the general GRE validation study, stratification of the samples on the Advanced Test on the basis of norms derived from the National Program further ensures a closer approximation to the population seeking admission to graduate education. In addition, the present studies use the ultimate culturally-defined criterion of successful completion of graduate education: the attainment of the doctorate degree.

II. GENERAL VALIDATION OF THE GRADUATE RECORD EXAMINATIONS AGAINST DOCTORATE ATTAINMENT

Purpose of the Study and Normative Basis for Defining Validation Samples

This part of the present report is concerned with the first of the two studies constituting the main content of the report, as described on page 2. Since the purpose of this first study is to validate the Graduate Record Examinations (GRE) on a broad-based sample of applicants for admission to graduate education in the sciences, it was necessary to select subjects from available applicants for NSF Graduate fellowships by stratification on available norms. Since two sets of test norms are generally available, an understanding of the purposes and sampling bases of these norms is essential to understanding the choice made for the stratification in the present study. It is also essential to understand the basis of conversion of raw test scores into the scaled scores widely used to report GRE performance.

... Scaling Technique

Conversions of raw test scores to the scaled scores are based on the performance of 2095 seniors tested in 11 institutions in 1952. These scales assign a mean of 500 and standard deviation of 100 on the Verbal and Quantitative tests, across all subjects in this group regardless of the field of Advanced Test taken. These statistics therefore vary among fields. The scales for the Advanced Tests were adjusted by regression methods to reflect differences in Aptitude Test scores among groups taking different Advanced Tests. The procedures used were developed by Dr. Ledyard R. Tucker and are reported in more detail by Schultz and Angoff(10).

... Two Norm Groups

Normative distributions on these scales are available from two previously cited sources(5,6). The norms for "basic reference groups" are based on performance of 3035 seniors in 21 institutions; the testing period is not stated but is probably 1956-7. More recently, Harvey and Lannholm have published normative data based on the performance of more than 25,000 students tested in the National Program during the 1960-1961 period. Although the number of institutions represented is not stated, it is surely larger than that reported

in the earlier norms. The test publishers properly disclaim that either set of norms represents "national norms", in spite of wide geographic and institutional representation. However, it is these norms that are in fact available and in use.

A comparison of the two sets of norms, those for "basic reference groups", and those from the 1960-1961 National Program, shows the latter group to be superior. It is also more explicitly relevant to the question of graduate school selection, and based on more recent information obtained in much larger samples. For these reasons the Harvey-Lannholm norms, rather than those for "basic reference groups" were used as the basis for defining stratified samples for the present validation study.

... National Program Norms Used

Starting with approximately 4500 former fellowship applicants having complete GRE data, validation groups were defined by stratification on the norms for the Advanced Test taken. This was done within field of specialization as defined by the field of Advanced Test. The 1954-1957 applicants for first-year fellowships and those intermediate applicants with Advanced Test scaled scores below 500 were distributed on Advanced Test score within field of Advanced Test and sex. Stratification was performed by dropping at random sufficient numbers of cases within 20-point scaled score intervals on the Advanced Test so that the remaining group would match the 1960-1961 National Program distributions reported by Harvey and Lannholm.

... Aptitude Test Not Used in Stratification

This stratification on the Advanced Test does not necessarily stratify the samples with equal rigor with respect to either the Verbal or Quantitative test norms. However, the inter-test correlations assure reasonably adequate coverage of the range of talent on Verbal and Quantitative tests within field of Advanced Test taken. In view of other factors preventing perfect control of the validation samples, it was decided not to attempt multivariate stratification, but to base both stratification and definition of field of study on the Advanced Test. Also with respect to the Aptitude Test, norms are based on subjects from <u>all</u> fields included, not just the science fields for which data in the present study are available.

... Sex Composition of Samples

Due to the fact that there are marked sex differences in doctorate attainment rates, it was decided to define separate validation samples for men and women, wherever sufficient cases were available to do so. However, this decision posed certain dilemmas in connection with the stratification on the National Program norms. Although differential norms by sex for groups including non-scientists are provided for the Aptitude Test, an unspecified mixture of sexes was used in norming the Advanced Tests. Female applicants for fellowships are available in quantities necessary for separate treatment in biology and chemistry. Insufficient numbers of cases were found in the other fields for forming separate samples. In psychology, in spite of a favorable sex balance, the small total number of cases available argued for a mixed sample, which is justified because of the larger proportion of females typical in this field. In the remaining fields--engineering, geology, mathematics, and physics--only samples of male subjects were used, stratified on the appropriate test norms. Both the male and female samples in biology and chemistry were stratified on the mixed norms in each field; failure to do so under these conditions might well bias validities by taking advantage of differential doctorate attainment rates by sex rather than by ability. In spite of these various compromises with ideal stratification and control of the validation samples, it is believed that the resulting samples are meaningful and relevant to the use of the validation data in generalizing to the broad-based groups seeking admission to graduate education.

... The Stanine Scale

In order to facilitate data processing, the reporting of distributions resulting from stratification, and the reporting of doctorate attainment rates at various levels of measured ability, the GRE scaled scores were converted after stratification to a nine-category normalized scale. The resulting scale is called the stanine scale and the scores resulting from this conversion are called stanines. A copy of the stanine conversion tables used is presented in Appendix A of this report. These tables relate the scaled scores and the derived stanines so that the equivalent categories of scaled scores used can be recovered in interpreting reported information.

Table 1

Sample Size, Means, Standard Deviations, and Percentage Distributions of Advanced Test Stanines, By Field and Sex

Validation Sample	Sample Size	Stanine on Advanced Test 1	2	3	4	5	6	7	8	9	Mean	Standard Deviation
Theoretical	---	4.0	7.0	11.0	17.0	20.0	17.0	11.0	7.0	4.0	5.00	1.96
All Cases Combined	2488	3.7	6.8	12.9	15.7	19.8	17.2	11.8	7.4	4.6	5.05	1.98
Biology												
Male	320	3.1	7.5	11.9	15.9	20.0	18.4	12.8	5.3	5.0	5.05	1.96
Female	140	5.0	4.3	12.1	17.1	22.1	16.4	10.7	8.6	3.6	5.04	1.94
Total	460	3.7	6.5	12.0	16.3	20.7	17.8	12.2	6.3	4.6	5.05	1.94
Chemistry												
Male	500	3.8	7.2	12.8	18.0	18.6	16.6	11.0	8.0	4.0	4.98	1.99
Female	160	3.1	6.9	16.9	13.8	17.5	16.3	14.4	8.1	3.1	5.01	1.99
Total	660	3.6	7.1	13.8	17.0	18.3	16.5	11.8	8.0	3.8	4.99	1.98
Engineering												
Male	300	2.0	6.0	15.0	13.0	22.0	19.0	12.0	7.0	4.0	5.11	1.88
Geology												
Male	119	4.2	5.9	13.4	16.8	20.2	14.3	14.3	7.6	3.4	5.01	1.96
Mathematics												
Male	250	4.8	5.6	12.0	17.6	20.0	15.2	12.8	4.0	8.0	5.07	2.07
Physics												
Male	600	4.3	7.5	12.0	14.0	19.8	17.3	11.0	9.0	5.0	5.09	2.04
Psychology												
Total	99	2.0	9.1	13.1	16.2	17.2	22.2	8.1	9.1	3.0	5.00	1.93

... Test Score Distributions

As a check on the stratification and stanine conversion operations, the percentage distributions, means, and standard deviations for the Advanced Test stanines were obtained for each field-by-sex sample. These data are presented with the number of cases in each stratified sample in Table 1. In addition, the theoretical stanine percentage distribution, which in fact defines the stanine scale, is presented in

the top row of the table as a basis for evaluating the outcome of the operations. The distributions are shown graphically in Figs. 1 and 2. With minor exceptions, not expected to bias validity data, the match to the Harvey-Lannholm norms on the Advanced Test is quite adequate.

... 95% Were First-year Applicants

Of the 2488 subjects in the combined stratified samples, 2374 or 95.4% were first-year applicants. As it turned out, using four years of applicants as a base for the stratification, it was not necessary to include intermediate applicants to develop the lower portions of the test distributions. With the exception of psychology where there were 86.9% first-year applicants, all fields had over 90% first-year applicants. However, biology had only 91%. Since biology and psychology are the two fields with appreciable proportions of female applicants, it is more difficult to fill in the lower portions of the distributions in female samples with first-year applicants only. The proportion of first-year applicants in the total stratified group is fairly uniform for all years of application, approximately one-fourth from each of the four program years, 1954-1957.

... Means on Aptitude Tests

The effects of stratification on the Advanced Test and of stanine conversion are shown in Table 2 which gives the Verbal and Quantitative score distributions in the validation samples. This table also presents the corresponding statistics, as available, for the normative groups as reported by Harvey and Lannholm(6). Statistics are presented for the validation samples in terms of both stanines and scaled scores, including those for the Advanced Test on which stratification was based.

The norms statistics on the Aptitude Test given in Table 2 show the familiar field differences. Means of the validation samples show similar field differences but more markedly than in the corresponding normative groups. Within any field, the standard deviations in the validation samples match those of the normative samples quite well; the means, however, are systematically higher in the validation samples, more so for the Quantitative than for the Verbal Test. Various reasons may be given for deviations of the validation samples from the within-field norms: sampling errors, imperfections of stratification, regression effects, and lack of control of sex composition. The crucial point in validation, however,

Figure 1

Stanine Score Distributions of Biology and Chemistry Males and Females

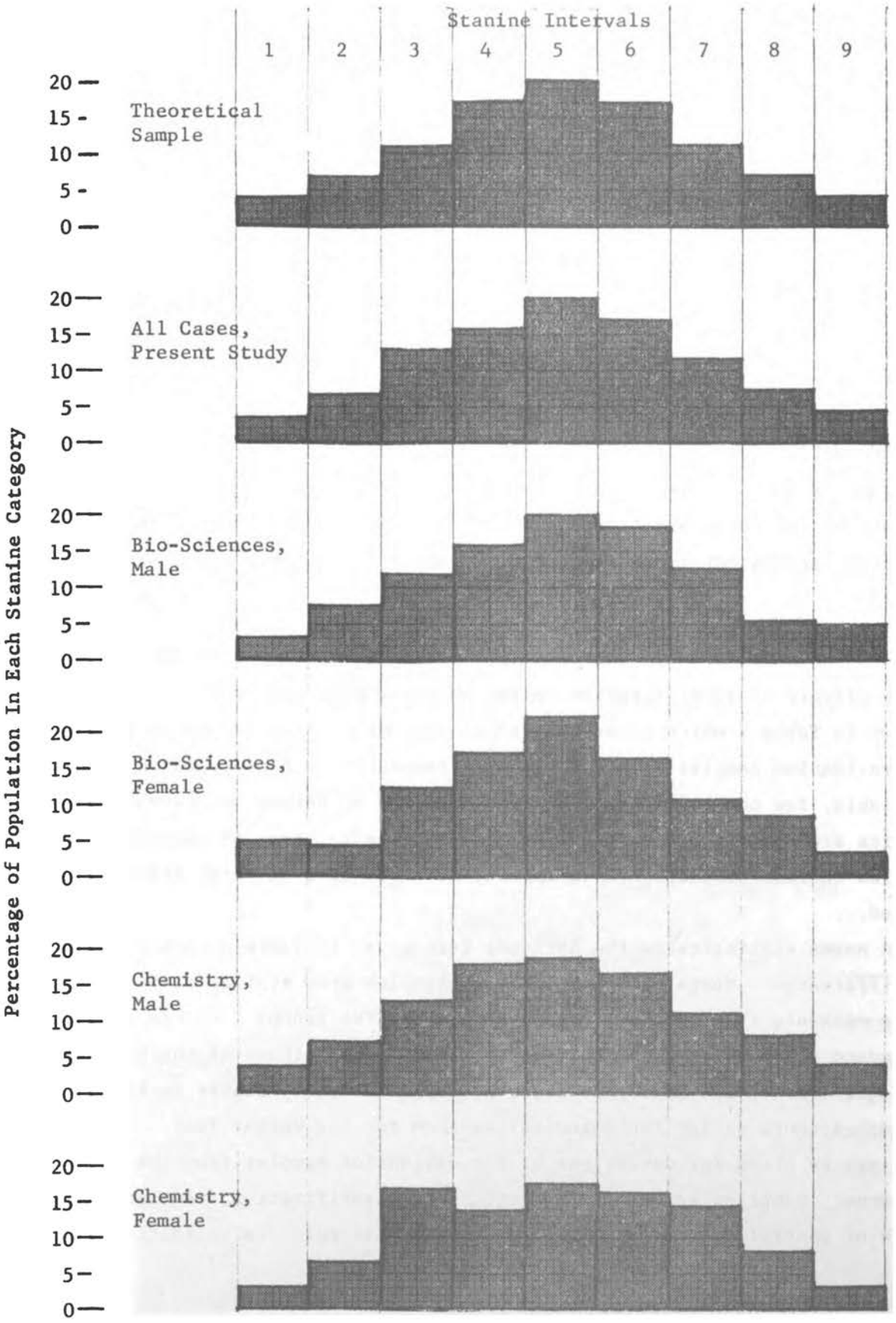


Figure 2

Stanine Distributions in Engineering, Geology, Math, Physics, and Psychology

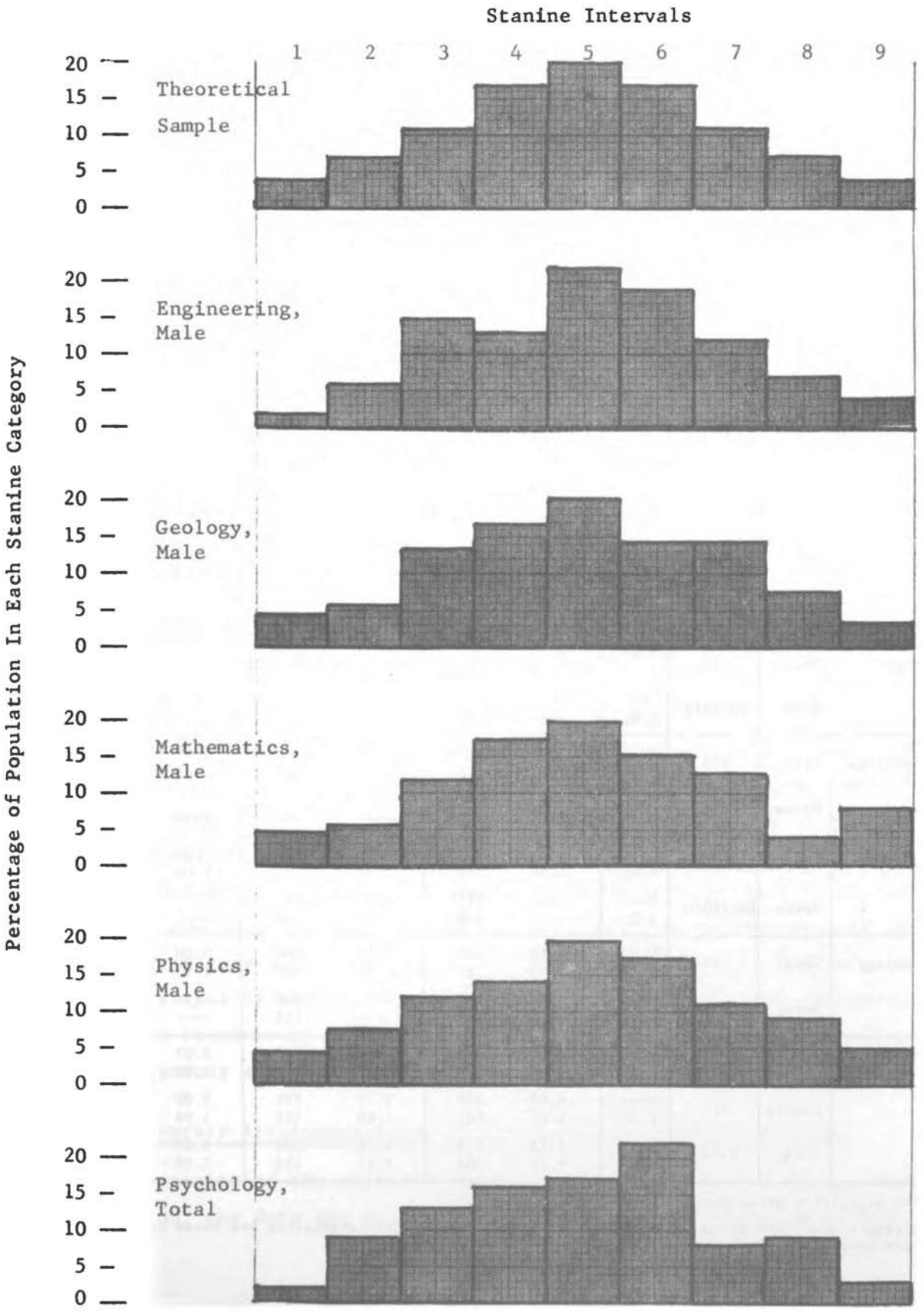


Table 2

Sample Size, Means and Standard Deviations of the Graduate Record Examinations By Field and Sex in Validation and Normative Samples[a]

Field	Sample	Sample Size[b]	Stat-istic	Verbal Test		Quantitative Test		Advanced Test	
				Stanine	Scaled Score	Stanine	Scaled Score	Stanine	Scaled Score
Biology	Male	320	Mean	5.21	547	5.70	579	5.05	578
			S.D.	1.87	112	1.73	117	1.96	120
	Female	140	Mean	5.80	585	5.19	543	5.04	578
			S.D.	1.65	99	1.54	102	1.94	117
	Total	460	Mean	5.39	558	5.54	568	5.05	578
			S.D.	1.82	110	1.70	114	1.94	119
	Norms	830/910	Mean	----	529	----	528	----	577
			S.D.	----	109	----	105	----	120
Chemistry	Male	500	Mean	5.33	555	6.68	646	4.98	596
			S.D.	1.90	112	1.61	111	1.99	120
	Female	160	Mean	6.14	607	6.36	622	5.01	593
			S.D.	1.76	104	1.65	109	1.99	113
	Total	660	Mean	5.53	567	6.61	640	4.99	595
			S.D.	1.89	113	1.60	110	1.98	118
	Norms	588/686	Mean	----	541	----	614	----	592
			S.D.	----	116	----	110	----	113
Engineering	Male	300	Mean	5.14	541	7.47	701	5.11	626
			S.D.	1.55	95	1.28	85	1.88	114
	Norms	887/977	Mean	----	531	----	682	----	615
			S.D.	----	110	----	94	----	107
Geology	Male	119	Mean	5.55	569	6.47	628	5.01	621
			S.D.	1.63	97	1.60	109	1.96	103
	Norms	307/370	Mean	----	535	----	575	----	617
			S.D.	----	101	----	105	----	101
Mathematics	Male	250	Mean	5.89	587	7.30	690	5.07	589
			S.D.	1.97	118	1.46	103	2.07	157
	Norms	680/824	Mean	----	577	----	677	----	581
			S.D.	----	117	----	102	----	146
Physics	Male	600	Mean	6.37	617	7.53	706	5.09	567
			S.D.	1.59	96	1.31	91	2.04	121
	Norms	886/1070	Mean	----	592	----	684	----	563
			S.D.	----	108	----	89	----	116
Psychology	Total	99	Mean	6.26	613	5.51	568	5.00	552
			S.D.	1.43	87	1.52	103	1.93	91
	Norms	1490/1682	Mean	----	570	----	528	----	550
			S.D.	----	102	----	115	----	93
Total	Male	2161	Mean	5.68	575	6.91	662	5.05	589
			S.D.	1.82	109	1.62	112	1.99	125
	Female	327	Mean	6.02	599	5.76	581	5.02	584
			S.D.	1.67	101	1.68	112	1.96	113
	Total	2488	Mean	5.73	578	6.76	652	5.05	588
			S.D.	1.79	108	1.67	116	1.98	123

[a] Norms statistics after Harvey and Lannholm Tables 3-5.

[b] Two sample sizes are given for normative samples; Advanced Test statistics are based on the larger sample, Aptitude Test statistics on the smaller sample.

is whether these effects would seriously bias the validities. Since the standard deviations are well matched to those of the norm groups and the means generally less than a quarter of a standard deviation higher, the validities are not likely to be seriously affected. However, in biology and psychology the distributions on the Quantitiative Test were found to be more markedly skewed than those in the other fields. These distributions are not shown here.

Relations Between Field of Advanced Test Taken and Field of Doctorate Attained

Up to this point the discussion of the general GRE validation study has been focussed on the definition and characteristics of the validation samples. Earlier, most of the crucial questions regarding the meaning of doctorate attainment criteria were discussed. Before presenting validity data, however, it is relevant to ascertain whether those persons who attained a doctorate did so in the same field in which their Advanced Test was taken. This is important, not only because of different doctorate attainment rates in different fields, but also because the field definition of the validation samples is based on the field of Advanced Test taken.

... Field-switching is Rare

Table 3 shows for each field of Advanced Test taken the percentage of doctorate recipients taking that test who achieved their doctorate in each field. Thus, in the first column, of the 140 taking the Biology test, 97.8% earned doctorates in the biological and agricultural sciences, 0.7% (1 person) in the medical sciences, and 1.4% in psychology. Other fields are to be interpreted similarly. The percentages remaining within field of the Advanced Test taken are larger in the validation group than the percentages of science doctorates in general who remain in the same field from baccalaureate to doctorate stages of education. For the purpose of the present validation study, this is no loss; in fact it makes the doctorate attainment criterion more homogeneous and clear-cut in meaning.

Doctorate Attainment Rates as a Function of GRE Score Levels

The results of the validity study will be presented in various forms. In this section, the data are presented in the form of doctorate attainment rates

Table 3

Field of Doctorate for Those Taking Various Advanced Tests

Field of PhD	Advanced Test Taken						
	Biology	Chemistry	Engineering	Geology	Mathematics	Physics	Psychology
Medical Sciences[a]	0.7	1.0	0.0	0.0	0.0	0.0	7.7
Biological & Agricultural Sci.[b]	97.8	14.0	0.0	0.0	3.3	3.8	0.0
Chemistry	0.0	79.0	1.9	0.0	1.6	0.5	0.0
Engineering	0.0	2.1	86.8	0.0	1.6	2.8	0.0
Earth Sciences	0.0	1.0	1.9	90.3	0.0	0.5	0.0
Mathematics	0.0	0.0	1.9	0.0	77.0	2.8	0.0
Physics	0.0	1.7	7.5	0.0	8.2	89.1	0.0
Psychology & Anthropology	1.4	0.3	0.0	6.5	1.6	0.5	88.5
Social Sciences	0.0	0.0	0.0	3.2	0.0	0.0	0.0
Arts & Humanities	0.0	0.3	0.0	0.0	4.9	0.0	0.0
Education	0.0	0.0	0.0	0.0	1.6	0.0	3.8
Professional Fields	0.0	0.3	0.0	0.0	0.0	0.0	0.0
Total Fields	100.0	100.0	100.0	100.0	100.0	100.0	100.0
No. of PhD's	140	287	53	31	61	211	26

[a] includes clinical psychology doctorates.

[b] includes doctorates in biometrics, biochemistry, and biophysics.

at various GRE score levels. The attainment rates used are those defined in terms of the August 1964 cutoff data and represent attainment within 10 years of baccalaureate completion. At a given score level, or in a category defined by a range of scores, the reported attainment rates are estimates of the probability that one performing at that level on the GRE will complete doctorate education.

Doctorate attainment rates by level of Advanced Test score are presented in Table 4. For each of the nine validation samples, the number of cases, number of doctorates, and percentage of doctorates are shown in terms of stanines grouped into five categories: stanines 1-2, 3-4, 5, 6-7, and 8-9. This places about 11% of the cases in the end categories, 29% in the intermediate categories, and 20% in the middle category. The data are also presented in the last column for the total sample.

Line graphs of the relation of doctorate attainment rate to Advanced Test score level are presented in Figure 3. In those samples where instability was observed under the 5-category stanine grouping, the stanines were regrouped into high (stanines 7-9), medium (stanines 4-6), and low (stanines 1-3) categories and are so plotted. In the sample of male subjects taking the Advanced Test in Chemistry, the combination of total field rate, sample size, and validity permits a more detailed picture in terms of all nine stanine levels. To render the curves comparable between samples, the differences in grouping of stanines from one sample to another were taken into account in plotting. The number of plotted points shown on a given curve reveals which stanine grouping was used (3,5, or 9 categories). Interpolation and extrapolation must be done, if at all, with considerable caution and with awareness of the small number of cases determining many of the points. In both Table 4 and Figure 3, the stanine categories may be translated back to scaled score categories by consulting the conversion tables presented in Appendix A of this report.

... Validity Slopes Are Positive

Except for instabilities noted previously, all curves show a generally steady increase in the probability of doctorate attainment as the test performance level increases. This is consistent with the fundamental meaning of the validity of a test against a given criterion. It does not mean, of course, perfect validity; some with low scores complete a doctorate and some with high scores do not. Nevertheless, the generally positive slopes of these curves indicate that the tests are useful in reducing the guesswork in predicting who will complete graduate education.

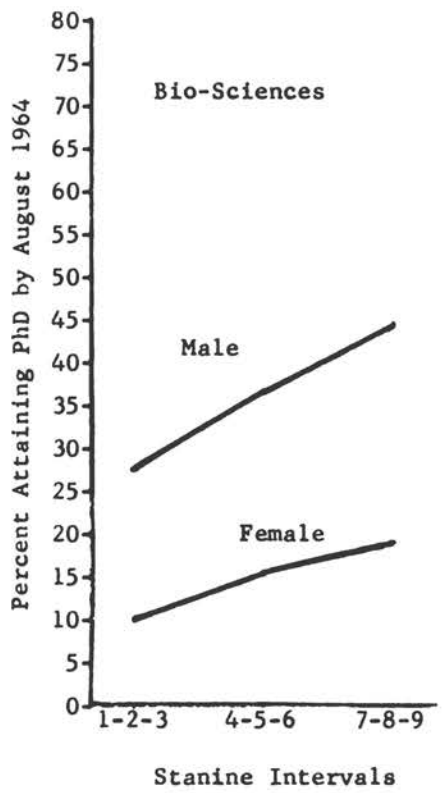
Table 4

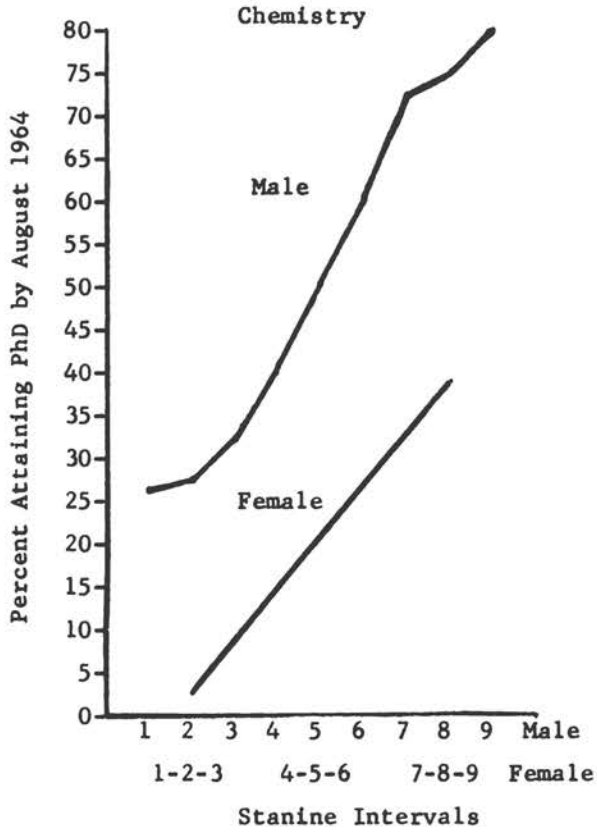
Doctorate Attainment Rate As A Function of Performance Level On The Advanced Test

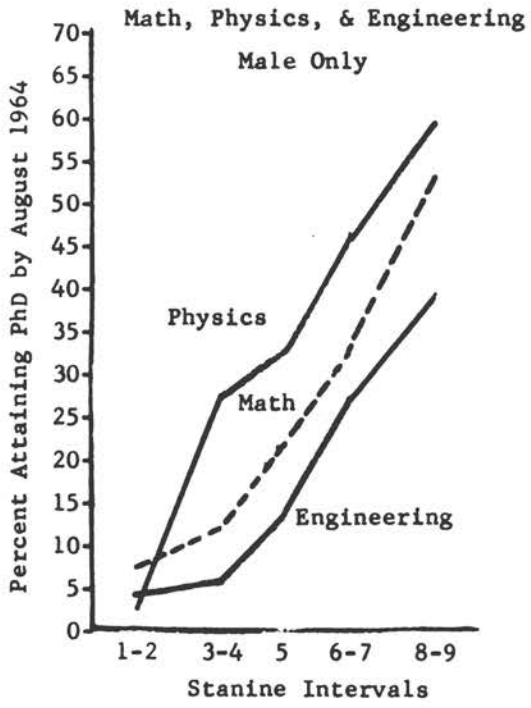
Sample	Statistic	PhD Attainment in Stanine Group 1-2	3-4	5	6-7	8-9	Total Sample
Biology Male	No. of Cases	34	89	64	100	33	320
	No. of PhD's	7	28	28	39	15	117
	% PhD	20.6	31.5	43.7	39.0	45.5	36.6
Biology Female	No. of Cases	13	41	25	38	17	140
	No. of PhD's	0	5	4	8	5	23
	% PhD	0.0	12.2	16.1	21.0	29.4	16.4
Chemistry Male	No. of Cases	55	154	93	138	60	500
	No. of PhD's	15	58	46	90	46	256
	% PhD	27.3	37.7	50.5	65.2	76.7	51.2
Chemistry Female	No. of Cases	16	49	26	49	18	160
	No. of PhD's	0	5	5	15	10	31
	% PhD	0.0	10.2	3.6	30.6	55.6	19.4
Engineering Male	No. of Cases	24	84	66	93	33	300
	No. of PhD's	1	5	9	25	13	53
	% PhD	4.2	6.0	13.6	26.9	39.4	17.7
Geology Male	No. of Cases	12	36	24	34	13	119
	No. of PhD's	3	8	4	8	8	31
	% PhD	25.0	22.2	16.7	23.5	61.5	26.1
Mathematics Male	No. of Cases	26	74	50	70	30	250
	No. of PhD's	2	9	11	23	16	61
	% PhD	7.7	12.2	22.0	32.9	53.3	24.4
Physics Male	No. of Cases	71	156	119	170	84	600
	No. of PhD's	2	43	39	79	50	213
	% PhD	2.8	27.6	32.8	46.5	59.5	35.5
Psychology Total	No. of Cases	11	29	17	30	12	99
	No. of PhD's	0	6	4	10	6	26
	% PhD	0.0	20.7	23.5	33.3	50.0	26.3

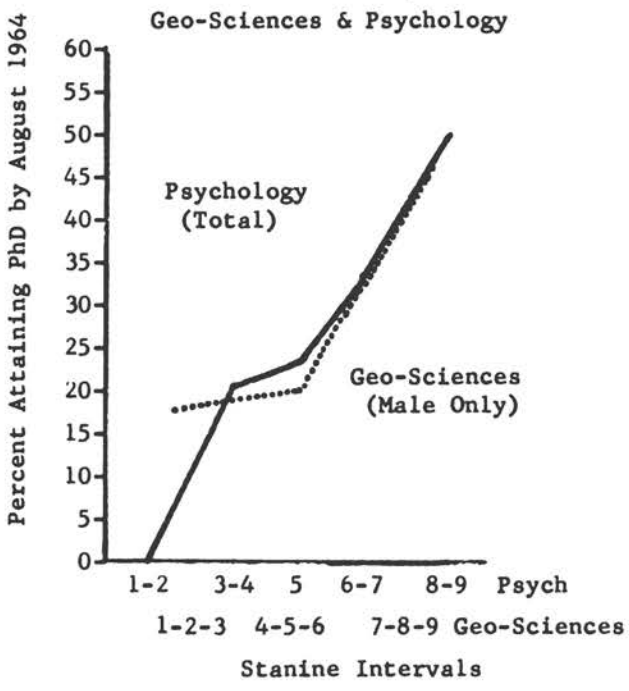
Figure 3

Doctorate Attainment as A Function of Advanced Test Stanine









... The Aptitude Tests Relate to PhD Attainment, Too

Doctorate attainment rates by level of Verbal Test score are presented in Table 5; similar data for the Quantitative Test are presented in Table 6. Although the five stanine groupings are defined as was done with the Advanced Test, the stanine conversions were made using the among-fields norms with the resulting skewness within fields, discussed in a previous section. This means that the numbers of cases on which attainment rates in the lower categories are based are smaller and therefore the rates are subject to larger sampling errors. In the case of the Quantitative Test the more "quantitative" science fields may be devoid of cases in the "low" category. In such a case the presumed doctorate attainment rate, while not computable, is zero because no people at this level applied for graduate study in that field.

... Validity Needs Quantitative Statement

Although less marked than in the case of the Advanced Test, these data again show the trend of higher doctorate attainment rates at higher test performance levels. The trend is somewhat stronger for the Quantitative than for the Verbal Test.

In order for doctorate attainment rate data to be maximally useful for interpretation of scores on the Graduate Record Examinations, such rates should be established on considerably larger samples. This recommendation and related suggestions are discussed more fully in the final part of this report. Meanwhile, it is instructive to evaluate these data on a basis which permits more ready comparison with validation results in other studies. This method of presenting validation data involves validity coefficients, which represent the degree of correlation between the test and the criterion of accomplishment. The following section presents these validation statistics for the Graduate Record Examinations against the various criteria of doctorate attainment.

Table 5

Doctorate Attainment Rate As A Function of Performance Level On The Verbal Test

Sample	Statistic	PhD Attainment in Stanine Group 1-2	3-4	5	6-7	8-9	Total Sample
Biology Male	No. of Cases	23	90	68	100	39	320
	No. of PhD's	3	28	23	40	23	117
	% PhD	15.0	31.1	33.8	40.0	59.0	36.6
Biology Female	No. of Cases	2	28	24	65	20	140
	No. of PhD's	0	3	4	13	3	23
	% PhD	0.0	10.7	16.0	20.0	15.0	16.4
Chemistry Male	No. of Cases	28	141	98	164	69	500
	No. of PhD's	11	62	52	90	41	256
	% PhD	39.3	44.0	53.1	54.9	59.4	51.2
Chemistry Female	No. of Cases	6	19	26	73	36	160
	No. of PhD's	0	3	5	12	11	31
	% PhD	0.0	15.8	19.2	16.4	30.6	19.4
Engineering Male	No. of Cases	4	114	68	88	26	300
	No. of PhD's	0	8	13	20	12	53
	% PhD	0.0	7.0	19.1	22.7	46.2	17.7
Geology Male	No. of Cases	4	26	27	47	15	119
	No. of PhD's	0	2	8	13	8	31
	% PhD	0.0	7.7	29.6	27.7	53.3	26.1
Mathematics Male	No. of Cases	11	46	54	77	62	250
	No. of PhD's	0	3	11	27	20	61
	% PhD	0.0	6.5	20.4	35.1	32.3	24.4
Physics Male	No. of Cases	4	74	101	261	160	600
	No. of PhD's	0	16	25	154	65	213
	% PhD	0.0	21.6	24.8	41.0	40.6	35.5
Psychology Total	No. of Cases	1	9	19	53	17	99
	No. of PhD's	0	1	4	14	7	26
	% PhD	0.0	11.1	21.1	26.4	41.2	26.3

Table 6

Doctorate Attainment Rate As A Function of Performance Level On The Quantitative Test

Sample	Statistic	PhD Attainment in Stanine Group 1-2	3-4	5	6-7	8-9	Total Sample
Biology Male	No. of Cases	16	61	61	135	47	320
	No. of PhD's	1	22	18	47	29	117
	% PhD	6.3	36.1	29.5	34.8	61.7	36.6
Biology Female	No. of Cases	8	39	30	54	9	140
	No. of PhD's	1	4	3	13	2	23
	% PhD	12.5	10.3	10.0	24.1	22.2	16.4
Chemistry Male	No. of Cases	2	49	59	235	155	500
	No. of PhD's	0	13	29	117	97	256
	% PhD	0.0	26.5	49.2	49.8	62.6	51.2
Chemistry Female	No. of Cases	4	17	25	69	45	160
	No. of PhD's	0	1	3	12	15	31
	% PhD	0.0	5.9	12.0	17.4	33.3	19.4
Engineering Male	No. of Cases	0	4	15	132	149	300
	No. of PhD's	0	0	0	20	33	53
	% PhD	(0.0)	0.0	0.0	15.1	22.1	17.7
Geology Male	No. of Cases	0	12	25	51	31	119
	No. of PhD's	0	2	2	13	14	31
	% PhD	(0.0)	16.7	8.0	25.5	45.2	26.1
Mathematics Male	No. of Cases	0	9	21	95	125	250
	No. of PhD's	0	0	3	12	46	61
	% PhD	(0.0)	0.0	14.3	12.6	36.8	24.4
Physics Male	No. of Cases	0	13	28	232	327	600
	No. of PhD's	0	0	5	57	151	213
	% PhD	(0.0)	0.0	17.9	24.6	46.2	35.5
Psychology Total	No. of Cases	2	21	33	31	12	99
	No. of PhD's	0	4	9	8	5	26
	% PhD	0.0	19.0	27.3	25.8	41.7	26.3

Validation Statistics for the GRE Against Doctorate Attainment Criteria

Correlation coefficients for the Graduate Record Examinations, individually and in combination, were computed against the three doctorate attainment criteria: (1) gross time lapse from baccalaureate, coded as shown on page 4, (2) the dichotomous attainment criterion defined as attainment by the August 1964 cutoff date, and (3) a second dichotomous criterion defined by cutting the time lapse distribution in each field just below the average value for that field.

... A Point About Biserial Coefficients

The dichotomous criteria may be considered as true dichotomies, i.e., one either does or does not complete a doctorate within the defined limit. If one takes this view, the appropriate correlation coefficient is the point-biserial coefficient. On the other hand, a serious case can be made for treating these discrete categories as mere dichotomizations of an underlying continuum. Not only are there degrees of successful approach to doctorate attainment (early attrition, terminal Master's degree, and "all but dissertation" stages), but there are also variations in the time taken to complete the doctorate. Where such an underlying continuum may be assumed to be normally distributed, the biserial rather than the point-biserial coefficient may be used. Actually the underlying continuum here is BA-PhD time lapse, the distribution of which is known to be skewed. To permit the reader to choose his assumptions, both coefficients will be presented in this section.

... Continuous Variables Have An Advantage

It should be noted that coefficients based on dichotomous criteria are subject to larger sampling errors than those for continuous criteria, and that the errors are larger in samples where attainment rates deviate markedly from 50%. The coefficients obtained on the coded time lapse criterion are probably the most trustworthy. They are probably about 5% low due to the effect of grouping non-doctorates into a single end category of the coded variable.

... Validities Are Positive

Validities are presented in Table 7 for each of the GRE tests and for the optimally weighted composite of the three tests as determined by multiple correlation methods. Within each validation sample and for each predictor, validity coefficients are reported separately for the three criteria, and in the case of the dichotomous criteria, separately by type of coefficient (point-biserial or biserial). In the prediction of the continuous time lapse criterion, the obtained validity coefficients were all negative--that is, the higher the score, the shorter the time lapse. By reversing the sign of this variable, the tests may be seen to be positive as predictors of the speed of doctorate attainment. This has been done in Table 7 by calling the criterion "reflected time lapse" and reversing the signs on the validity coefficients to positive.

... Choose Your Criterion

In examining Table 7 it is recommended that validities be compared within a given row, representing a particular predictor and sample. Such a comparison bears on the question of which criterion is most predictable. Since the three criteria are different renderings of the same basic information, they are very highly correlated. The validities in a given row are therefore essentially the same within sampling errors of the coefficients, and excepting the bias inherent in biserial validities when the normality assumption is not met. If there is any advantage to one criterion over either of the others, it is in favor of the reflected time lapse measure.

... Advanced Test Most Valid

Examination of the validities within a sample and for a given criterion shows that the Advanced Test validities are the highest among those for the individual tests in all but two samples: the sample of males taking the Biology test and the total (male) sample taking the Geology test. Usually the differences between the Advanced Test validities and those for either the Verbal or Quantitative Test are greater than can be accounted for by the fact that the samples were stratified on the Advanced Test, with the result that the distributions on the Verbal and Quantitative tests are somewhat skewed. The exceptions in biology and geology are primarily due to random fluctuations in the coefficients.

Table 7

Validities of the Graduate Record Examinations Against Doctorate Attainment Criteria [a]

Sample	No. of Cases	PhD Attainment Rate	Predictor	Criteria [b]				
				Reflected Time Lapse	PhD I Pbis	PhD I Bis	PhD II Pbis	PhD II Bis
Biology Male	320	36.6	Verbal	23	20	26	20	26
			Quantitative	23	21	27	21	27
			Advanced	18	14	18	17	22
			Composite	26	23	29	23	30
Biology Female	140	16.4	Verbal	14	06	09	14	22
			Quantitative	20	11	17	22	35
			Advanced	23	17	26	23	37
			Composite	25	19	29	26	41
Chemistry Male	500	51.2	Verbal	16	12	15	13	16
			Quantitative	26	21	26	22	28
			Advanced	38	31	39	34	43
			Composite	39	33	41	35	44
Chemistry Female	160	19.4	Verbal	15	17	25	17	25
			Quantitative	29	27	39	27	39
			Advanced	38	37	37	37	54
			Composite	40	38	38	38	55
Engineering Male	300	17.7	Verbal	28	28	41	28	42
			Quantitative	21	21	31	19	28
			Advanced	32	31	45	31	46
			Composite	35	34	50	34	50
Geology Male	119	26.1	Verbal	25	30	41	26	36
			Quantitative	26	27	37	24	33
			Advanced	22	20	27	22	30
			Composite	31	33	45	32	44
Mathematics Male	250	24.4	Verbal	21	22	30	19	27
			Quantitative	25	26	36	23	32
			Advanced	36	34	47	34	48
			Composite	36	35	48	34	48
Physics Male	600	35.5	Verbal	16	15	19	12	16
			Quantitative	26	26	33	23	30
			Advanced	34	32	41	30	39
			Composite	35	33	43	32	42
Psychology Total	99	26.3	Verbal	13	13	18	17	24
			Quantitative	17	13	18	16	22
			Advanced	33	25	34	30	42
			Composite	34	25	34	30	42

[a] Decimal points have been omitted from the validity coefficients.

[b] PhD I = attainment by August 1964; PhD II = attainment by field median. See p. 25

... But Validities Are Modest

While these validities are not high, they are uniformly positive and statistically significant. The tests are therefore useful in appraising the ability of a student applying for admission to graduate education. The order of magnitude of these coefficients indicates, however, that factors other than the abilities measured by these tests are significant in doctorate attainment.

... The Three Predictors as a Team

Can better prediction of doctorate attainment by means of the Graduate Record Examinations be obtained by combining the scores from all three tests than by using any one alone? To answer this question the optimum composite of the three tests was defined separately for each sample and the validity of the composite determined by multiple correlation methods. The resulting multiple validities are presented in Table 7 in the rows labeled "composite" in the predictor column. Here it may be seen that combining the three test scores increases the validity only slightly, but consistently above that obtained for the Advanced Test alone. Within a given sample the increase is within random sampling fluctuations. Further information on this matter is presented in Table 8 where the intercorrelations among the three tests and the standard partial regression weights are reported.

... Why So Little Gain?

The intercorrelations obtained in these samples are higher than those reported in "basic reference groups". These intercorrelations, which enter into the computations of the composites, measure redundancy of information among the three tests; if this redundancy is low, the combining of the independent information from the three tests is more effective than if it is high. The regression weights define the relative contribution of each test in defining the optimum composite for that sample. As expected from the pattern of validities for the individual tests, the regression weights show the Advanced Test usually carrying most of the weight in multiple prediction. In most fields the Quantitative Test makes a small contribution, except in those fields where there is a high quantitative component in the Advanced Test. The contribution of the Verbal Test is highly variable from sample to sample. In several samples it adds nothing beyond the information in the Advanced and Quantitative Tests. In some

Table 8

Correlations Among GRE Tests and Standard Regression Weights Against PhD Attainment[a]

Sample	No. of Cases	GRE Test	Intercorrelations		Regression Weights		
			Quantitative	Advanced	Reflected Time Lapse	PhD I	PhD II
Biology Male	320	Verbal	61	71	15	12	08
		Quantitative	--	51	14	14	14
		Advanced	--	--	00	-02	04
Biology Female	140	Verbal	65	69	-08	-12	-11
		Quantitative	--	65	12	03	15
		Advanced	--	--	21	24	21
Chemistry Male	500	Verbal	64	53	-11	-10	-11
		Quantitative	--	58	12	10	10
		Advanced	--	--	37	31	33
Chemistry Female	160	Verbal	59	47	-13	-07	-07
		Quantitative	--	54	19	12	12
		Advanced	--	--	33	34	34
Engineering Male	300	Verbal	52	52	15	15	31
		Quantitative	--	46	02	03	00
		Advanced	--	--	23	22	23
Geology Male	119	Verbal	53	37	13	19	16
		Quantitative	--	40	14	14	11
		Advanced	--	--	11	08	11
Mathematics Male	250	Verbal	57	59	-02	00	-02
		Quantitative	--	64	04	06	03
		Advanced	--	--	34	30	33
Physics Male	600	Verbal	52	51	-05	-06	-07
		Quantitative	--	49	09	13	11
		Advanced	--	--	31	27	28
Psychology Total	99	Verbal	26	40	00	04	06
		Quantitative	--	49	00	00	02
		Advanced	--	--	32	24	27

[a] Decimal points have been omitted from the coefficients and weights.

samples it has a positive weight; in others it has a negative weight, presumably where the Advanced Test already contains sufficient verbal content for prediction of these criteria.

... Curvilinear Regression Tested

Examination of some of the curves shown in Figure 3 suggests the possibility that the relationship between test score and doctorate attainment might be non-linear. Departures from a rectilinear relationship are frequently found in actual data plots, but can most often be explained simply on the basis of random sampling variations which reverse themselves or disappear when the study is replicated. It is possible, however, to test whether a given relationship is curvilinear in its fundamental nature by using a computational formula that allows for the existence of such relationships and produces a better-fitting regression line if such a fundamental relationship is actually present. This formula takes into account not only the additive relationships, but also the multiplicative relationships between variables. In the terms of mathematical statistics, it is the complete second degree polynomial regression model. The multiple correlation computed by this statistical model will be higher than that of the standard linear regression formula whenever the data depart from linear relationships either through chance variations or because some fundamental curvilinearity is present. In the present case, use of the formula which included these second degree terms added two to three points to the multiple correlation, on the average. This was not deemed statistically significant, but due to random sampling fluctuations only, as shown by formulas which allow for shrinkage because of random errors in the data. Thus the usual linear model is seen to exhaust the predictive significance of the data in this case; no real improvement in prediction of doctorate attainment can be achieved by taking into account the multiplicative relationships among the three GRE tests or between the tests and the criterion of doctorate attainment.

III. PREDICTION OF DOCTORATE ATTAINMENT IN THE FOLLOW-UP GROUP

Plan of the Study

... A Different Context

In contrast to the study reported in Part II, this study was carried out in the context of a follow-up of the 1955 and 1956 first-year and intermediate applicants for National Science Foundation graduate fellowships. In the first report on the follow-up study, Technical Report #24, the 3623 subjects were characterized in terms of response rate, field of academic study, year and level of first application, award status, and doctorate attainment. In the present study, doctorate attainment information is used as a criterion for the validation of the selection instruments for both questionnaire respondents and non-respondents without regard to award status. For this purpose, subjects with complete information on control and selection variables were used, reducing the number of subjects to 3491.

Field of study is here defined in terms of stated major field at time of fellowship application, which determines the evaluating panel to which an application is assigned. Ninety-one percent of the subjects in the study took the Advanced Test in the same panel field. Another four percent were Biology panel subjects who took the Chemistry test and were assumed to be biochemists.

... Additional Predictors

In addition to validation of the Graduate Record Examinations in the follow-up group, this study examines validities of the undergraduate science-mathematics grade-point average, the reference report average rating, and the extent to which these variables added to the validity of the GRE. The grade-point averages were not available in the earlier terminal follow-up study, so this was the first opportunity to examine the effectiveness of the grade-point average in the selection of National Science Foundation fellowship applicants. It is reasonable to expect the grades and reference report ratings to add something, be it achievement motivation, studiousness, or academic interest, to the abilities measured by the Graduate Record Examinations in prediction of doctorate attainment criteria. To oversimplify, it is plausible to claim that the Verbal and Quantitative Tests

measure aptitudes developed at earlier stages of learning, that the Advanced Test measures information content and its usage specific to a given field of study, and that the remaining, non-GRE variables tap motivational, social, and interest factors in academic achievement. If this is so, it is reasonable to expect the Advanced Test to add to the prediction of doctorate attainment beyond what can be predicted from the Aptitude Test, and the results of the study reported in Part II confirm this supposition. It is also reasonable to expect the grade-point average and reference report ratings to add still further prediction. This will be tested in the present study, by comparing two optimum composites: one which does and one which does not include these additional variables.

Another interesting question that can be answered by the present study is how optimal composites developed in this study for predicting doctorate attainment compare with the composite of the same variables, the summary score, designed to maximize agreement with the panel ratings, and how both composites compare with the quality group ratings by the panels. The quality group includes panel use of non-quantitative and moderating information not included in the composites; insofar as such information is valid and properly used by the panels, it should add prediction beyond the quantitative information included in the various composites.

... Description of Samples

For this study, the follow-up group was sorted by field, level, and sex, and then combined into samples large enough for reasonably stable validity estimates while maintaining as much homogeneity as possible. Seventeen groups were so defined. They are described in Tables 9 and 10, pp. 34 and 35. Table 9 gives means and standard deviations for the test data, and Table 10 the same statistics for undergraduate grade-point average, reference report ratings, summary score, and quality group determinations. Due to the inclusion of a larger proportion of intermediate applicants, the doctorate attainment rates are somewhat higher in these samples than in those of the first study.

Validation Results

...Single and Composite Predictors

In all 17 samples, validity was measured against the criterion of attainment of the doctorate by August 1964, less than 10 years from the baccalaureate degree. The point-biserial coefficients are reported (rather than biserial coefficients)

because of the marked departures from normality on both predictors and criterion in these samples. In nine of the samples, validities were also computed against the time lapse criterion. Where necessary, the signs of the coefficients were reflected so that a good score is always positively related to high achievement. This was necessary for quality group, where the highest group is "1", the lowest, "6", and for time lapse (short lapse is preferred). The validities of the various predictors are presented in the columns of Table 11. For a given sample, the validities are first presented for the Verbal, Quantitative, and Advanced Tests, then for the grade-point average and reference report rating. Four multiple validity correlations are then presented: (1) for a composite of the three GRE tests, (2) for a composite of the GRE tests plus the grade-point average, (3) for a composite of the GRE tests plus reference report average rating, and finally (4) for a composite of all five selection variables. The validities of the summary score and quality group complete the set. For the record, the intercorrelations among the five selection variables are presented in Appendix B.

... Rank the Validities

Examination of the first five columns of validities in Table 11 shows those for the Advanced Test to be generally highest, followed by those for the Quantitative Test; the validities for grade-point average and reference report variables are next, and about equal; the Verbal Test is generally lowest in validity. This pattern of relative validities is fairly typical of the prediction of academic criteria, although there are variations also according to which criterion is chosen: the time lapse to the doctorate or the percentage reaching the doctorate by August 1964. There is variation also by level; the coefficients are lower in the intermediate level applicants than in the first-year applicants, apparently because academic attrition and different factors of self-selection tend to restrict the variance on the intermediate group. When the validities of the various measures are compared for the two criteria, the grade-point average and reference report variables show up with higher coefficients in prediction of time lapse than of eventual doctorate completion, in contrast to the test variables, which show no systematic difference between the two criteria. It may be that the grade-point average and reference reports include something of a drive or motivational element that is absent in the test scores, and that this element is related to acceleration, but not to eventual degree attainment.

Table 9

Distribution Statistics on Graduate Record Examinations in Follow-up Groups

Field	Sex	Level	No. of Cases	% PhD	Verbal Test Mean	Verbal Test Std. Dev.	Quantitative Test Mean	Quantitative Test Std. Dev.	Advanced Test Mean	Advanced Test Std. Dev.
Biology	Male	1	204	59	574	112	605	118	606	120
	Male	2	316	78	582	103	618	116	665	111
	Male	1+2	520	71	579	107	613	117	642	118
	Female	1+2	191	32	603	99	569	101	618	119
Chemistry	Male	1	354	71	584	104	689	104	679	120
	Male	2	309	88	605	93	694	89	782	82
	Male	1+2	663	79	594	100	691	97	727	116
	Female	1+2	84	26	609	112	642	106	630	138
Engineering	Male	1	423	36	575	95	731	77	688	113
	Male	2	162	62	600	87	750	68	726	114
	Male	1+2	585	43	582	93	736	75	699	114
Geology	Male + Female	1+2	176	58	594	94	645	106	666	117
Mathematics	Male	1+2	264	56	628	117	725	102	688	158
Physics	Male	1	497	56	627	93	727	84	613	140
	Male	2	272	78	645	90	745	70	724	118
	Male	1+2	769	64	634	92	733	80	652	142
Psychology	Male + Female	1+2	161	63	655	88	619	113	625	113

Table 10

Distribution Statistics on Grade Average, Reference Ratings, Summary Score, Quality Group

Field	Sex	Level	No. of Cases	% PhD	Grade-point Average Mean	Grade-point Average Std. Dev.	Reference Ratings Mean	Reference Ratings Std. Dev.	Summary Score Mean	Summary Score Std. Dev.	Quality Group Mean	Quality Group Std. Dev.
Biology	Male	1	204	59	2.3	0.4	4.2	0.9	59.5	12.0	3.5	1.5
	Male	2	316	78	2.1	0.5	3.9	0.8	60.0	10.7	3.5	1.4
	Male	1+2	520	71	2.2	0.5	4.0	0.9	59.8	11.2	3.5	1.4
	Female	1+2	191	32	2.4	0.4	4.0	0.9	59.5	11.3	3.6	1.5
Chemistry	Male	1	354	71	2.5	0.5	4.2	1.0	64.9	13.0	3.3	1.3
	Male	2	309	88	2.3	0.4	3.9	0.9	67.6	9.4	3.4	1.1
	Male	1+2	663	79	2.4	0.5	4.0	1.0	66.1	11.6	3.3	1.2
	Female	1+2	84	26	2.4	0.4	3.9	1.0	61.0	13.2	3.7	1.3
Engineering	Male	1	423	36	2.4	0.4	4.2	0.8	65.8	10.3	3.2	1.5
	Male	2	162	62	2.4	0.4	4.0	0.8	67.1	10.4	3.4	1.3
	Male	1+2	585	43	2.4	0.4	4.2	0.8	66.2	10.4	3.3	1.4
Geology	Male + Female	1+2	176	58	2.3	0.4	4.2	0.8	63.0	10.3	3.3	1.4
Mathematics	Male	1+2	264	56	2.5	0.4	4.2	0.9	67.0	13.2	3.4	1.5
Physics	Male	1	497	56	2.5	0.4	4.1	0.9	63.5	12.8	3.4	1.6
	Male	2	272	78	2.5	0.4	3.9	0.8	67.4	9.7	3.6	1.3
	Male	1+2	769	64	2.5	0.4	4.0	0.9	64.9	11.9	3.5	1.5
Psychology	Male + Female	1+2	161	63	2.3	0.4	4.2	0.8	62.2	10.3	3.7	1.7

... Combine the Predictors

Examination of the four columns of coefficients under the general heading of "Composites" shows what gain can be achieved in prediction by combining one or more of the predictor variables. For the present criteria, it appears that the Advanced Test usually does just about as well as does the three-test composite, although in the more verbal fields such as biology and psychology the Verbal Test does add a little to the prediction. Beyond this, the addition of grade-point average and reference report ratings makes little difference, and when one is added, no further gain is made by adding the other. It is important to note that these findings apply strictly to the academic criteria used in the present study, and that generalization to other criteria is unwarranted. For example, it has been found in the study of terminal level applicants that, for the prediction of on-the-job ratings of scientific competence and productivity, the reference report ratings were the best predictors.

... Compare the Combinations

The final two columns in Table 11 present correlations of summary score and quality group with doctorate attainment and time lapse criteria. The comparison of these two columns with each other and with the optimally-weighted composite of "all predictors" is of interest. It should be noted at the beginning that the summary score is a composite of the same five predictor variables, but one optimally-weighted for prediction of the quality group decision of the panels. Also, it is a single composite designed to be optimal for a wide range of fields and levels, with men and women combined, whereas the "all predictor" composite is optimal for the particular group and criterion of each row. It should also be noted that the quality group is unique in this set of composites in that it is a judgment of a panel of scientists as to "ability" for scientific work. It is made in the light of the evidence from the several quantitative variables and also all other evidence in the individual's application materials, including such allowances as seem proper for the school at which the grade-point average was earned, and known or suspected tendency of the particular reference reporters to be "hard raters" or "easy raters". It is a decision which is intended to encompass not only ability to complete the doctorate, but also to be effective later as a practicing scientist. Still, it is a subjective judgment, and subject to the limitations always present in such judgments.

Table 11

Doctorate Attainment Validities of Selection Variables and Composites in Follow-up Groups

Field	Sex	Level	No. of Cases	% PhD	Criterion[a]	Selection Variables					Composites					
						V	Q	A	GPA	RRA	Total GRE	GRE+ GPA	GRE+ RRA	All Selectors	SS[b]	QG[c]
Biology	M	1	204	59	PhD	28	28	28	25	19	32	35	33	35	33	34
	M	2	316	78	PhD	05	13	10	14	13	15	17	18	19	17	20
	M	1+2	520	71	PhD	16	20	22	13	12	25	25	25	26	24	26
	M	1+2	520	71	T.L.	21	27	25	30	23	30	37	34	38	36	38
	F	1+2	191	32	PhD	18	21	33	13	08	35	35	35	36	27	33
	F	1+2	191	32	T.L.	23	22	26	16	13	29	30	29	30	28	36
Chemistry	M	1	354	71	PhD	19	26	44	30	29	45	47	46	47	41	39
	M	2	309	88	PhD	06	13	23	09	10	24	24	24	24	20	20
	M	1+2	663	79	PhD	16	22	42	18	18	43	43	43	43	36	31
	M	1+2	663	79	T.L.	11	22	34	29	24	36	41	39	41	36	39
	F	1+2	84	26	PhD	07	29	49	22	13	55	55	55	55	37	35
	F	1+2	84	26	T.L.	10	27	46	21	11	50	50	50	50	35	34
Engineering	M	1	423	36	PhD	07	16	21	20	21	22	26	27	27	26	25
	M	2	162	62	PhD	10	15	19	27	17	19	29	23	29	25	21
	M	1+2	585	43	PhD	11	18	23	20	17	24	27	26	27	26	22
	M	1+2	585	43	T.L.	11	19	24	23	22	25	30	30	31	30	27
Geology	M+F	1+2	176	58	PhD	27	30	44	11	21	47	47	47	47	43	29
	M+F	1+2	176	58	T.L.	22	25	37	15	33	39	40	45	46	46	35
Mathematics	M	1+2	264	56	PhD	30	21	39	15	20	41	41	41	41	40	39
	M	1+2	264	56	T.L.	26	28	41	26	27	41	42	43	44	43	44
Physics	M	1	497	56	PhD	21	27	39	32	22	39	43	40	44	38	36
	M	2	272	78	PhD	05	02	14	08	10	15	16	17	17	15	18
	M	1+2	769	64	PhD	18	22	38	22	15	38	39	38	39	35	29
	M	1+2	769	64	T.L.	15	26	37	30	23	38	42	39	42	39	38
Psychology	M+F	1+2	161	63	PhD	28	19	20	14	23	28	31	32	33	30	28
	M+F	1+2	161	63	T.L.	16	20	25	21	23	26	32	31	35	33	31

[a] PhD=PhD attainment; T.L.=Time Lapse (reflected) b. SS = Summary Score. See text for explanation

[c] Signs of validities reversed to reflect Quality Group variable

Because the summary score and "all selector" composites are made up of the same five variables, but with different weights, it is impossible for the summary score to have the higher validity of the two, as the "all selector" composite, by definition, has optimal weights in these samples for prediction of the present criterion. It is possible, however, for quality group to be more valid, as it contains additional elements which may have validity. On the other hand, it suffers known decrement from one source and possible decrement from another. It is expressed in only six steps; this coarse grouping or crudity of measurement restricts and attenuates the possible validity by perhaps four or five percent. It may also have elements of unreliability due to its subjective nature--that is, the effect of human judgment may be to introduce more noise than signal. As it turns out, the summary score is only slightly less valid than the "all selector" composite, the difference ranging from 1 to 7 correlation points for the men and from 3 to 18 points for the women. The greatest difference is for the group of 84 female chemists. This is attributable chiefly to the relatively small size of this group, with resultant capitalization in the "all selector" composite on larger errors of random sampling in the computation of the multiple correlation. This results in an inflated validity for the "all selector" composite. The summary score is, in effect, a cross-validated composite in which such error-capitalization is absent. The quality group validities are lower than the summary score validities in 17 of the comparisons, better in 9, and equal in one case. In three of the nine times when quality group predicts better than summary score, it also predicts better than the "all selector" composite--twice in biology where it is always better than summary score, and once in physics. On the average, however, quality group is about 1.3 correlation points lower than summary score in predicting the doctorate attainment and time lapse criteria. This decrement is just about what is expected because of the limitation to six predictor categories. Only in geology is the difference in favor of the summary score greater than 10 points, for reasons which are not apparent from the data.

... Does Human Judgment Add to Prediction of PhD Attainment?

The fact that the "cross-validation" correlations described above are almost as good as the multiple correlation derived from the same data deserves some attention. In particular, the moderating effect of human judgment in quality

group determination is noteworthy. The judgmental operation quite evidently adds nothing to the mechanical formula of the summary score in predicting this criterion. Whether it adds to the prediction of an on-the-job criterion remains for investigation in a separate study. Perhaps improved predictions are made in certain types of cases by the addition of a human judgmental process. If so, the improvement is offset in the present instance by decreased accuracy in prediction in other cases, so that the average effect is no loss and no gain. It should be emphatically noted here that the function of the panels in reviewing the applications is not completely defined by the doctoral attainment criteria used in this validity investigation, and that the findings with regard to predictive validity of quality group should not be generalized to other criteria as may later develop on the job.

... How Reliable are PhD Attainment Criteria?

The general level of the validity coefficients reported in Table 11 is not high, even for those variables, or composites of those variables, for which reliabilities are known to be quite high. Some of the possible reasons for these modest validities, and a more thorough consideration of the question of reliability and validity of the several variables, are discussed in Appendix C. Suffice it here to note that some of the many factors involved in completion of the doctorate are inherent in the individual, and of these, a few have been measured. Others, such as factors concerned with motivation, are very weakly and indirectly measured in the predictors available for the present study. Other factors, outside the individual and beyond his control, inhere in the institutions, the departments, and the individual professors with whom the students come in contact. It is quite conceivable that improved prediction of these criteria could be obtained by either or both of two routes: more adequate measurements of individual factors of interest, motivation, and commitment to graduate study, and more adequate measurements of institutional and departmental factors. The latter would be used to moderate the criterion, to present a criterion measure in which each individual with the same ability would have more nearly an even chance for the same criterion score. If this could be done, then the composite of all individual factors that could come under the heading of a broadly defined "ability" could be more accurately evaluated. There is suggestive evidence that such improvement could be made; in Table 11 the pattern of validities in the bio-sciences group is quite different

from that of the other fields. The data as they stand, however, give only faint clues as to how such improvement in prediction could be effected.

... Attrition of Able Students Disturbing

An important manpower consideration is also highlighted by the validity statistics here presented. It is evident that the attrition rates at high ability levels as measured by the Graduate Record Examinations are disturbingly high. Validity of the more inclusive composites shows the same picture: a gradual but marked attrition, particularly in the first or second year of graduate study, of very able students. It would seem that some special kind of academic motivation is required to defer the gratification of a good income in affluent times, or to postpone marriage and family life more than a few years beyond the early 20's, for people who have already had 16 years of academic environment and may be yearning for a change. In effect, the panelists in the present instance, or those who pass on graduate school admissions in the larger context to which the present data might be applied, are faced with a problem of great complexity and difficulty. They are being asked to make judgments with regard to these motivational factors as well as others, among highly selected students, for a criterion that is itself hard to define.

IV. SUMMARY AND CONCLUSIONS

... Two Samples, Two Sets of Predictors

Two inter-related studies are here reported. The first employed a sample of fellowship applicants so selected as to resemble as closely as possible the general population of applicants for graduate school, and examined the validity of the Graduate Record Examinations for prediction of time lapse between the baccalaureate and doctorate degrees, and probability of attaining the doctorate within a defined time period. The second study employed larger samples of applicants for National Science Foundation fellowships in several fields, which samples were selected for a follow-up study. These groups were higher in ability than those in the first study. For the second study, in addition to the Graduate Record Examinations, data were available on the validity of undergraduate grade-point average and the average ratings of professors, submitted in reference reports. The increments to valid prediction attainable by use of these variables in combination with the test scores were examined, as was the validity of the quality group decision of the panels which evaluate fellowship applicants.

... Main Results Summarized

Both studies confirm that the Graduate Record Examinations have significant validity for prediction of doctorate-attainment criteria, even within the upper levels of test performance. Both studies confirm the greater validity of the Advanced Test over the Aptitude Test. When the three tests are used in an optimal combination, a small increment in validity is usually, but not always, attained. The second study shows in addition that moderate additional validity may be obtained by combining into a single formula the undergraduate grade-point average and the ratings of professors, providing this weighting is optimized by the multiple correlation technique. It is evident from both studies that factors above and beyond these variables are involved. It is probable that individual factors of drive, interest, and motivation are involved, and that additional institutional, departmental, and other environmental factors would have to be considered in attaining a higher prediction.

... Criterion Definition Vital

The nature of the criterion here employed is crucial in evaluation of the significance of the present findings. Many students seek graduate school entrance with the master's degree as the goal. The present study does not speak to the validity of the tests or other measures for this purpose, as practically all the students involved were candidates for the doctorate degree, and the attainment of this degree, or the time taken to attain it, were the criteria. The criterion here used is also very different from those employed in the on-the-job follow-up studies of fellowship candidates. For such on-the-job criteria, it has been found in other studies that the reference reports are the best predictors. It may be said, however, in connection with the doctorate-attainment criteria, that the doctorate degree is becoming an almost universal prerequisite to leadership in the science fields here involved. In the present studies, it has been found that the speed of attaining the doctorate degree is probably the most useful criterion, although criteria described in terms of the percentage of a group attaining the doctorate by a specified time may also be valuable. The chief value of the doctorate-attainment percentage criterion measure is its capability of specifying the probability of eventual doctorate attainment for a given level of test performance.

... The Matter of Human Judgment

An additional finding is worthy of particular note because of its significance for prediction and selection purposes, and because it is subject to the possibility of misunderstanding and over-generalization. It was found that a composite of test scores, grades, and professors' ratings which is optimal for prediction of the quality group decision of the evaluation panels is almost as good for predicting doctorate attainment as is a composite specially weighted for this purpose. Furthermore, either composite is, for most fields, a better predictor of doctorate attainment than is quality group itself, with its subjective weighting of these variables as well as others in the application materials. The subjective judgment element adds as much noise as signal for prediction of this criterion. It must not be assumed, however, that this means that the panels can be superseded by a statistical formula, as it is not the prediction of doctorate attainment which the panels are called upon to make. They are required to make an over-all judgment of "ability"--a far broader and more inclusive judgment than that involved only in attaining the doctorate degree. They also perform a number of other valuable functions, including communications, which cannot be evaluated statistically.

... Recommendations For Study and Action

A number of recommendations flow from consideration of the present findings. With regard to the Graduate Record Examinations, a much more searching validation could be performed in about three or four years, when much larger numbers of tested cases would have reached the doctorate level, or have dropped by the wayside. Further, a much better study could be made if the publisher would, at that time, make available more extensive norms based upon recently-tested cases sorted by field of Advanced Test, by sex, and by type of institution attended or region of the country involved. Such a study should employ samples of cases stratified on the basis of a composite of the Advanced Test and Aptitude Test, and much larger numbers of cases in order to better stabilize the validity statistics. Stabilization would be especially useful in the case of doctorate attainment rates at various ability levels. In addition, the probability of doctorate attainment at various periods of time lapse beyond the baccalaureate, for a given level of measured ability, could be computed. The general form of such

information is illustrated by the diagram on the inside front cover of this report. The results of such a more extensive and well-controlled study should be far more useful than any presently available, including the results of the present study, for purposes of guidance and graduate school selection.

REFERENCES

1. "Doctorate Production in United States Universities , 1920-1962", NAS-NRC Publication 1142, National Academy of Sciences-National Research Council, 1963.

2. "A Study of Graduate Fellowship Applicants in Terms of Ph.D. Attainment", Technical Report No. 18, by John A. Creager, National Academy of Sciences, 22 March 1961.

3. "Some Characteristics of First-Year and Intermediate Fellowship Applicants Eight to Ten Years Later", Technical Report No. 24, by John A. Creager, National Academy of Sciences, 16 August 1965.

4. Lannholm, Gerald V. The use of the Graduate Record Examinations in Appraising Graduate Study Candidates. Graduate Record Examinations Special Report 62-3, October, 1962. Princeton, N.J., Educational Testing Service.

5. Graduate Record Examination Scores for Basic Reference Groups. Third Printing, 1961. Princeton, N. J., Educational Testing Service.

6. Harvey, Philip R. and Lannholm, Gerald V. The Performance of Candidates Tested in the National Program For Graduate School Selection, 1960-1961. Graduate Record Examinations Special Report 62-1, January, 1962. Princeton, N.J., Educational Testing Service.

7. Lannholm, Gerald V. and Schrader, William B. Predicting Graduate School Success, An Evaluation of the Effectiveness of the GRE, Princeton, N.J., Educational Testing Service.

8. Lannholm, Gerald V. Abstracts of selected studies on the relationship between scores on the Graduate Record Examinations and graduate school performance. Graduate Record Examinations Special Report 60-3. November, 1960. Princeton, N.J., Educational Testing Service.

9. Tupes, Ernest C. and Dubois, Donald B. The educational achievement of Air Force officers. Technical Note, WADC-TN-58-68. November, 1958. Personnel Laboratory, Wright Air Development Center, ARDC, USAF, Lackland AFB, Texas.

10. Schultz, Margaret K. and Angoff, William H. The Development of New Scales for the Aptitude and Advanced Tests of the Graduate Record Examinations. Research Bulletin 54-15. May 25, 1954. Princeton, N.J., Educational Testing Service.

APPENDIX A

Stanine Conversion Table for Converting Scaled Scores[a] to Normative Stanines

Test	Stanine								
	1	2	3	4	5	6	7	8	9
Verbal	20-33	34-37	38-43	44-50	51-56	57-63	64-68	69-73	74-90
Quantitative	20-32	33-36	37-42	43-49	50-56	57-63	64-70	71-76	77-90
Advanced									
Biology	20-36	37-42	43-48	49-54	55-60	61-66	67-72	73-78	79-95
Chemistry	34-39	40-44	45-50	51-56	57-62	63-67	68-73	74-79	80-95
Engineering	28-43	44-48	49-53	54-58	59-64	65-69	70-75	76-81	82-95
Geology	29-43	44-48	49-54	55-59	60-65	66-69	70-74	75-77	78-95
Mathematics	32-37	38-41	42-46	47-52	53-60	61-68	69-78	79-86	87-95
Physics	34-37	38-42	43-47	48-52	53-58	59-64	65-70	71-77	78-95
Psychology	20-37	38-43	44-48	49-53	54-57	58-62	63-66	67-71	72-95

[a] Scaled scores are reported as three-digit scores with the third digit always zero; this final zero-digit has been dropped from the scaled scores in this table.

APPENDIX B

Intercorrelations Among Selection Variables in Nine Validation Samples

Sample	Variable	Quantitative Test	Advanced Test	Grade-point Average	Reference Rating Average
Biology Male	Verbal Test	57	64	16	19
	Quantitative	--	47	30	25
	Advanced Test	--	--	23	21
	Grade Average	--	--	--	46
Biology Female	Verbal Test	60	61	18	22
	Quantitative	--	46	16	29
	Advanced Test	--	--	20	24
	Grade Average	--	--	--	50
Chemistry Male	Verbal Test	61	50	12	18
	Quantitative	--	52	29	29
	Advanced Test	--	--	23	28
	Grade Average	--	--	--	61
Chemistry Female	Verbal Test	69	54	30	23
	Quantitative	--	66	44	26
	Advanced Test	--	--	41	27
	Grade Average	--	--	--	65
Engineering Male	Verbal Test	50	44	16	14
	Quantitative	--	55	28	24
	Advanced Test	--	--	32	27
	Grade Average	--	--	--	60
Geology Total	Verbal Test	46	27	11	18
	Quantitative	--	39	32	15
	Advanced Test	--	--	16	24
	Grade Average	--	--	--	32
Mathematics Male	Verbal Test	65	60	17	22
	Quantitative	--	65	27	24
	Advanced Test	--	--	38	35
	Grade Average	--	--	--	49
Physics Male	Verbal Test	51	52	09	13
	Quantitative	--	59	30	28
	Advanced Test	--	--	29	30
	Grade Average	--	--	--	58
Psychology Total	Verbal Test	54	55	07	28
	Quantitative	--	59	08	29
	Advanced Test	--	--	06	18
	Grade Average	--	--	--	34

APPENDIX C

... Reliability and Validity

In a study of this kind some consideration of the technical details of the prediction problem is unavoidable. This includes an examination of the concepts of reliability and validity, with particular reference to their application to the relationship of predictor measurements to the criteria of accomplishment. A variable is said to be highly reliable if, on repeated measurement, the same reading is obtained. Unreliability is introduced when the same instrument gives varying results on repeated application, as would a cloth tape-measure that stretches unevenly under varying tension, or a volt-meter with corroded terminals. Yet there is no question of validity--the tape-measure records length, not weight or temperature, and the volt-meter records voltage, not pressure or humidity or radiation. Validity concerns the extent to which any variable, however reliable or unreliable, yields measurements which are predictive of some other attribute. In human measurements, height and weight are correlated, and thus height could be said to have some validity as a predictor of weight. This validity is, however, far from perfect, and is improved very little by increasing the reliability of the measurement of height. We can improve on the prediction of weight by combining measurements of height and girth, and by statistical techniques we could determine the optimal combining proportions for these two measurements. Still, regardless of the reliability of the measurements of height and girth, we would have an imperfect composite predictor of weight. With respect to the latter, we may find that its measurement, too, is unreliable. If the scale upon which the subjects are weighed has rusty knife-edges, or imperfect springs, the measurements of weight may also be unreliable, and hence not perfectly predictable by any set of measurements of height, girth, specific gravity of the body, or any other conceivable measurements.

... Composite Measures

In all psychological measurements, and in all measurements of human achievement, these concepts of reliability and validity are involved. In all of the variables with which this report deals, the measurements are complex. The impact and application of unreliability to test scores has been extensively explored, and estimates of test reliability are generally available, and have been reported here for the Graduate Record Examinations. The unreliability of the other

measurements is less well-understood. The extent to which two reference reporters will agree in their ratings has been explored, and found to be rather modest. It is for this reason that a composite of the ratings of at least three reporters is sought whenever possible. On the average it has been found that a composite of four reference reporters will agree with another similar composite to the extent indicated by a correlation coefficient of about .70. This is not nearly as high as the tests, but high enough to be very useful, and to be the best predictor of later on-the-job success.

... Academic Achievement Measurements

When we come to a measure such as undergraduate grade-point average, or to a criterion measure such as time lapse between baccalaureate and doctorate degrees, the usual meanings of the term reliability are inapplicable. We cannot get repeated measurements--cannot put the students through the same courses twice to determine whether or not they will get the same grades; they cannot go through graduate school twice to determine how long it will take. And yet it is evident that these attainment measures cannot be perfectly reliable--cannot possibly be predicted by any conceivable set of measurements of the individuals themselves. This is because they depend in part on institutional factors outside the individual's control, including the characteristics of the university as a whole, the department in which he does his work in particular, and even upon the person who happens to become his major adviser--and to some extent on all the other faculty members whose courses he takes. The basic difficulty here is that, although it is possible to make a list of the factors making for unreliability of these attainment measures, it is not possible from data presently available, to achieve any measurement of the degree of unreliability associated with any factor or combination of factors.