GRE

GRADUATE RECORD EXAMINATIONS

THE CRITERION PROBLEM: WHAT MEASURE OF SUCCESS IN GRADUATE EDUCATION?

Rodney T. Hartnett
and
Warren W. Willingham

GRE Board Research Report GREB No. 77-4R
March 1979

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

ETS

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

The Criterion Problem:

What Measure of Success in Graduate Education?

Rodney T. Hartnett

and

Warren W. Willingham

GRE Board Research Report GREB No. 77-4R

Abstract

A wide variety of potential indicators of graduate student performance are reviewed. Based on a scrutiny of relevant research literature, experience with recent and current research projects, and conversations with graduate faculty members and administrators, the various indicators are considered in two ways. First, they are analyzed within the framework of the traditional "criterion problem," that is, with respect to their adequacy as criteria in predicting graduate school performance. In this case emphasis is given to problems with the criteria that make it difficult to draw valid inferences about the relationship between selection measures and performance measures. Second, the various indicators are considered as an important process of the graduate program. In this case, attention is given to their adequacy as procedures for the evaluation of student performance--e.g., their clarity, fairness, and usefulness as feedback to students.

The assessment of graduate student performance is complex, both conceptually and technically, and the overall system of evaluation seems reasonably fair and sound in the sense that the students earning degrees are undoubtedly those who are more competent and deserving. Viewed separately, however, each of the various indicators of success has numerous shortcomings. As a result, the measures normally available for validation studies, for evaluation of student performance, or for program evaluation are often not as good as they should or could be.

Various general observations about the status of student evaluation practices are made, and particular attention is drawn to the view that many evaluation practices seem to be characterized by ambiguity with regard to their basic purpose. Finally, suggestions are offered for how assessment practices might be improved.

The Nature and Importance of the Criterion Problem

In any educational program a primary question is how one defines the criteria of successful performance. The so-called "criterion problem" has always been an important issue in validating admissions tests; what constitutes success also has a critical bearing on the very conception of a program and its objectives. Various measures are used to mark success: grades, comprehensive examinations, departmental ratings, and so on. Such measures serve as a principal or at least partial basis for evaluating student progress, planning the curriculum, and evaluating program effectiveness.

Obviously the question of defining the criteria of success lies at the heart of the educational program--its operation as well as its goals. Nonetheless, there is limited literature on the problem as it applies to graduate study. In fact, Hirschberg and Itkin (1978) recently asserted, "...there has been practically no attempt whatsoever at a thorough theoretical criterion analysis of graduate school success" (p. 1085).

Notions of what constitutes successful student performance and how it ought to be measured naturally vary widely across institutions, disciplines, and types of programs. It is very much a responsibility of individual institutions and departments to wrestle with an issue so central to educational policy. It is also true that the criteria used in graduate education, and how those criteria are assessed, have a very important bearing on the nature of the admissions tests of the Graduate Record Examinations Board, their effectiveness, and the feasibility of improving their quality.

Because of these considerations the GRE Board requested an overview of the criterion problem as it applies to graduate education. This report is based upon a review of relevant research literature, experience with recent research projects of the Board, and conversations with graduate faculty and administrators on visits by the authors to ten universities*.

*The ten universities, are: California (Berkeley), California (Santa Cruz), Duke, Michigan, North Carolina (Chapel Hill), Pennsylvania, Princeton, Stanford, Temple, and Utah.

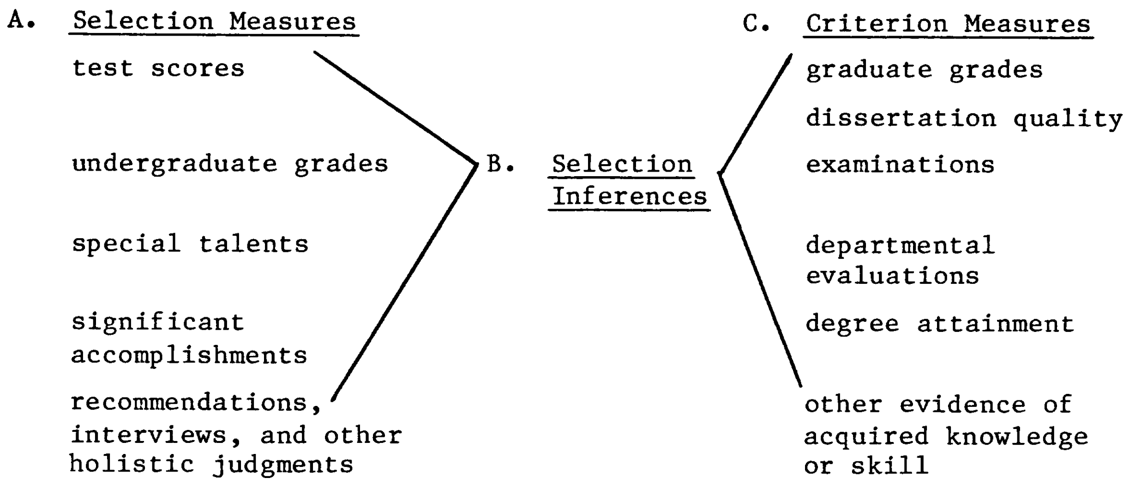
These consisted first of five visits undertaken in conjunction with a closely related project. At these universities, preliminary conversations were held with the senior academic officer about assessment of academic and nonacademic qualities of students at admissions and after enrollment. Typically graduate or undergraduate deans or directors of admissions participated in these discussions. In a second round of visits at five additional universities more intensive discussions with graduate deans and graduate faculty (from biology, history, and psychology) focused specifically on performance assessment--covering broad educational and philosophical issues as well as strengths and weaknesses of particular assessment practices.

On the basis of these conservations and our review of relevant research literature, we shall discuss the criterion question in two separate but related ways. First, we consider the importance of criteria for a better understanding of the selection problem, especially the adequacy of the criteria for making inferences about selection variables. Second, we consider the importance of criteria in their own right for the evaluation of student performance and for feedback to the student.

Subsequent sections are devoted to detailed discussion of three general types of criteria: 1) administrative criteria commonly used in graduate programs, 2) measures of professional accomplishment that are commonly available but not ordinarily used, and 3) measures that might be useful but have to be specially constructed. The purpose here is to summarize what is generally known about success criteria in graduate training and possible ways that the definition and use of such performance measures might be improved.

Success Criteria and the Selection Function

As the sponsor of examinations widely used in graduate admissions, it is the responsibility of the Graduate Record Examinations Board to insure that its tests offer a sound basis for selection. Since we assume that the use of tests results in the choice of students able to do well in graduate programs, a key issue is what constitutes doing well. Another way of posing the question is to ask: Within what framework one should examine the validity of an admissions test? Consider the following paradigm:



A. Selection Measures

test scores

undergraduate grades

special talents

significant accomplishments

recommendations, interviews, and other holistic judgments

B. Selection Inferences

C. Criterion Measures

graduate grades

dissertation quality

examinations

departmental evaluations

degree attainment

other evidence of acquired knowledge or skill

Strictly speaking, students are not selected only because of their standing on the selection measures in column A. These are largely historical and have only limited interest in and of themselves. We would like to use the measures under column C for selection but, of course, they lie in the future and can't be known at the time students are admitted. Students are actually selected on the basis of inferences about the relationship between selection measures and criterion measures. We infer that students will do well on the measures on column C if they have done well on the measures in column A.

This focus upon the underlying inference is not mere philosophic neatness. Notice what happens when we focus upon the validity of the inference rather than the validity of the selection test. In order for the inference to be valid we have to have well conceived measures on both sides. This view of the logic of test validity makes it clear that the GRE Board cannot be fully responsible for the validity of the examinations it sponsors without also concerning itself with the validity of the criterion measures.

There are two rather different but complementary ways in which the validity of selection inferences can be examined. One is predictive validity based upon statistical analysis of the relationship between selection measures and performance criteria; the other is construct validity based upon rational analysis and empirical studies.

Most of the literature on the validity of selection tests is concerned with predictive validity. At the graduate and professional level, validity coefficients are typically somewhat lower than at the undergraduate level (Willingham and Breland, 1977). There are several likely reasons for this typical finding (Willingham, 1976; Conrad, Trismen, and Miller, 1977; Wilson, 1978); a very important one is the restricted range of talent among students who enter most graduate programs. Closely associated with this range restriction is the problem of unreliable criterion measures due, for example, to the fact that graduate students often receive mostly A's with a few B's. Thus a search for empirical evidence of the validity of the selection inference may be hampered to a considerable extent by shortcomings in the criterion measure.

An alternate, rational approach to validation has long had professional sanction as an important element of construct validity (APA, 1974). This conception of validation has gained increasing favor in Supreme Court decisions, especially in the recent case of Washington vs. Davis (see also relevant discussions by Manning, 1978, and Lerner, 1978).

In supporting the validity of a selection test through rational analysis, one undertakes a careful examination of what traits and competences are assumed to be important in the training program. Such analysis might be further supported through special studies of the nature of performance in the educational program and research on what other abilities and competences seem related to that performance. Here too it is readily apparent that a narrow conception of the criterion of success in the educational program leads naturally to a narrow conception of relevant predictors, and possibly also to an erroneous conclusion that intrinsically valuable predictors lack demonstrable validity. These considerations argue further that a clear conception of success criteria is critical in considering possibly useful new measures of talent.

The Board has recently undertaken several projects aimed at the development of measures that would tap a broader variety of human talents and serve as a useful supplement to present examinations. These include projects to develop tests or inventories that measure scientific thinking (Frederiksen and Ward, 1978), cognitive style (Donlon, Reilly, and McKee, 1978), independent accomplishments (Baird, 1978), and analytic ability (Conrad, 1976). These projects are all reflections of the Board's plan since 1974 to provide the basis for broader interpretation of talent relevant to graduate training.

An important first step in the development of such measures is careful consideration of arguments and evidence supporting their use as selection measures. Demonstration of their empirical validity is equally important and clearly, such a demonstration requires adequate criteria. No matter how intrinsically sound and justifiable a selection measure may be on the basis of rational analysis, its empirical validation may founder on the shoals of an inadequate criterion. If new predictors do represent different types of talent important to success in graduate education, it is critical to provide an appropriate target for those predictors to shoot at; otherwise a validation effort may simply indicate that a valuable type of talent is unrelated to a graduate grade average that does not enjoy the confidence of the faculty anyway. Naturally any new GRE instrument must be specially evaluated as to its validity and technical soundness, but there is every reason to believe that an improved general understanding of the nature of success of graduate study should facilitate such work.

Success Criteria and the Educational Program

This report gives primary emphasis to those aspects of the criterion problem which are directly related to the selection function and the responsibilities of the Graduate Record Examinations Board. It is important to note, even if briefly, other important functions of criterion measures that bear directly upon the conception and character of the educational program.

The most obvious function of performance criteria is to provide a fair and relevant basis for student evaluation. The need for systematic evaluation of student progress has always aroused controversy in education and graduate study is no exception. The arguments are familiar. On the one hand, grades, examinations, and other routine hurdles are deemed necessary to motivate and direct students and to mark their progress. On the other hand, such evaluation is often perceived as a distraction from learning and an imposition of undesirable standardization on students pursuing somewhat different educational goals.

Those favoring evaluation usually win the argument on grounds of equity to students and the integrity of the program's content and standards. On the other hand, assessment of achievement in graduate programs is often highly individualized. Some would argue that there is too little clear specification of what is expected of students and inadequate evaluation of the extent to which students in a given program develop a useful common body of knowledge and skill.

A second useful function that criterion measures can serve is a basis for curriculum planning and program evaluation. It should be self evident that program design and evaluation depend upon agreement as to what students are expected to gain from the program. There is, however, considerable controversy in most fields as to what extent graduate training should focus upon a common curriculum. Usually some compromise is struck in the form of a core program with at least some common objectives shared by students in a given area. Devotion to curriculum specialization, individualized programs, and often individual faculty judgment of student performance make program evaluation exceedingly difficult.

Defining what constitutes success in a graduate program determines in the most fundamental way the very rationale of the program and what social function it serves. Success perceived in traditional terms of scholarship and command of the discipline is one thing; success perceived as the professional capability to fulfill a useful and needed social function may be something quite different. The implications of these differences understandably cause a good deal of tension and concern in many fields, especially with respect to important ancillary objectives of a department; e.g., matriculating larger numbers of those types of students previously underrepresented who may bring different traits and interests to graduate training, or the need to help graduate students find jobs at a time of reduced opportunity in many academic fields.

These various considerations emphasize the essential connections between the definition of "success" and the two vital functions: how students are selected and how they are educated. We now turn to the question of what specific issues deserve attention and how the general domain of success criteria might be usefully framed.

Framing the Issues

The previous discussion makes clear the importance of good definitions and accurate assessment of success. What specific issues deserve the attention of the GRE Board and graduate faculty generally? There are several classic criterion problems that apply to most selection situations in employment and education. These can be grouped conveniently as problems concerning reliability, intrinsic validity, and the range of success measures normally used. These classic problems apply to graduate education as well.

Reliability of criterion measures is an endemic problem partly because measures like ratings, grades, or completing the dissertation are usually developed for training or administrative purposes not necessarily to provide a fair, accurate, and consistent measure that discriminates among more and less successful individuals. Such measures sometimes make little distinction among people who are actually very different in competence or they make spurious distinctions that result more from artifactual aspects of the way the measures are collected for administrative purposes (as when virtually everyone passes an examination at the same level, or when variation in marks is due largely to habits of the graders). Routinely available criterion data also suffer a variety of practical problems that tend to reduce reliability of measures and stability of statistical results; e.g., small samples, highly selected groups, and criteria that are not compatible from one group to another.

Verifying the intrinsic validity of the criterion is another classic problem. Criteria are often complex and reflect a variety of skills, extraneous factors, and luck that have little to do with the real competence of most importance in the training program. Careful definition of the right signs of success is important not only because it influences who gets selected; this also helps to focus the educational effort in the right direction. A good illustration of a criterion lacking intrinsic validity is a written examination that is intended to assess practical, hands-on skill but is laden with verbal intelligence.

For example, even though a reading test may correlate highly with the final examination grade in a training course for a machine operator, the test may not be defensible for selection and may have a detrimental influence on the training program if reading is not an important part of the machine operator's actual job. Similarly, creative thinking may be an important part of a scientist's training and professional work, and a good test of creative thinking is not necessarily invalid if it is uncorrelated with a set of grades that largely reflect whether students complete designated assignments on schedule. Because of such considerations Gulliksen (1976) has urged persuasively that "all aptitude test development include evaluation of the criterion."

This point is related to a third type of deficiency frequently noted in criterion measures--the range of competencies included. Even if the criterion is a reasonably good reflection of successful overall performance, it may be that some extremely important but relatively rare type of competence in graduate study is overshadowed by the more usual types of performance normally reflected in grades. As a result, it may be almost impossible to find a substantial relationship between some fairly specific predictor and any global criterion of success, no matter how important the predictor may be to some types of graduate work or subsequent professional skill.

Various writers have argued that traditional measures used for graduate and undergraduate admissions place undue emphasis upon a strictly scholastic view of education; i.e., aptitude tests and grades do not recognize a diversity of purpose among disciplines and institutions, nor do they recognize the clear fact that the educational objectives of most faculty and most institutions are broader than pure academic competence (Wing and Wallach, 1971; Richards, Holland, & Lutz, 1967; Baird, 1976; Wallach, 1976).

In thinking about "success" it is useful to differentiate immediate, intermediate, and ultimate criteria (Thorndike, 1949).

Grades and end-of-year examination results are immediate criteria; career success is the ultimate criterion. Which is the more appropriate basis for judging the validity of selection measures? The purpose of selection tests is to predict success in the educational program, not success on the job. In Washington vs. Davis, the Supreme Court clearly endorsed the reasonableness of that purpose even in a training program for a specific job, but giving attention only to those academic measures (e.g., grades) that are readily available may foster an overly narrow view of training objectives. Training objectives in graduate education are defined with a long range view of career demands, but it is neither practical nor reasonable to validate tests against career success. Success often depends upon many personal factors that may have little to do with the types of competence graduate study is intended to develop. Furthermore, career success can be defined in many ways and good measures are extremely difficult to obtain. It is possible, however, to define *intermediate* criteria that focus upon competencies important in career success but which do not suffer the conceptual and practical disadvantages of "ultimate" measures. Many such competencies are reflected in "preprofessional" behavior in graduate school. Consider the following:

Program Objective	Pre-Professional Behavior
PRACTITIONER	--Demonstrated skill and interest in practical problems --Early involvement in professional affairs --Intern performance
TEACHER	--Teaching skills --Demonstrated interest and skill in helping students --Involvement in institutional affairs
SCHOLAR/SCIENTIST	--Independent accomplishments --Innovative work --Publications

Often, little systematic effort is made to evaluate such behavior in relation to training objectives; also, there is typically little effort to develop measures that predict these forms of success. Different training objectives are certainly not mutually exclusive across departments (see Clark, Hartnett, and Baird, 1976), though they do differ and may vary considerably in the case of particular specialities or individual students.

There are educational as well as social reasons for recognizing the full range of talent that may be called for in graduate study or subsequent professional work. Educators are familiar with both sides of the old argument concerning the extent to which graduate education should be directed toward professional training, but the changing market for individuals with advanced training in many fields compels even more serious consideration of objectives of graduate education and the concomitant implications for effective use of human talent. These concerns are all the more important considering that some groups of students who may be underrepresented in graduate training programs may have personal strengths that are quite useful in more professionally oriented graduate training programs though they may be less well prepared in those areas related to traditional academic scholarship.

One way of summarizing these issues is to ask to what extent it is both desirable and feasible to improve measures of success in graduate training with respect to the following uses of criteria measures:

--for the evaluation of student progress
--as a basis for examining the validity of selection measures
--as a basis for local self-study and program evaluation
--as a means of focusing on diverse talents and competencies relevant to graduate training and professional work
--as a means of focusing upon desired training objectives.

Three Classes of Criteria

In thinking about types of criteria that are used or might be used in graduate education, it is important to bear in mind the diversity of programs and the nature of the learning process in graduate programs. In many cases Master's programs are quite different from Ph.D. programs; programs differ considerably across fields and in the extent to which they are oriented toward professional training or scholarly work associated with the academic discipline. Another important consideration is the size and style of graduate departments. In many departments, particularly smaller ones, there is an apprentice tradition that places more emphasis upon intimate and informal appraisal that is tuned to the special interests and objectives of the individual student as opposed to more uniform expectations reflected in standard examination procedures. These variations can be dramatic--so much so that a reasonable way of measuring success of students in one department might be quite unreasonable in another. With these caveats in mind, it is useful for the purposes of this report to distinguish three broad classes of criterion measures as follows:

Administrative indices of academic competence--formal measures used to define the student's academic standing in the department

- --course grades
- --examinations (comprehensives, preliminaries, orals, language)
- --completion of dissertation
- --ratings (general evaluation, areas of competence, major assignments)
- --degree status (degree attained, progress toward degree, time to degree)

Evidence of professional accomplishment--information typically available but often not formally recorded as criteria of success

- --recognition (awards, honors, fellowships, appointments)
- --publications (articles, speeches, technical reports, important role in preparation of a book or major report)
- --professional activities (active in departmental affairs, participation in professional meetings, student leader, active speaker or organizer)
- --independent achievements (independent research, design equipment, significant discovery, important achievement)
- --field experience (intern ratings, work achievements)
- --teaching skill (student ratings, other available evidence)

Specially developed criteria--constructed measure of some important broad competence or special skill, particularly measures that can be scored objectively

- --artificial criteria (specially devised tests or simulalions requiring a critical skill that has high construct or face validity)
- --work samples (a product of the student's work, selected papers, precis, abstracts)
- --common examinations (standard essay questions, problems, or assignments)
- --ratings (a standard profile of important competencies and characteristics).

It is readily recognized that the first of these categories refers to those criterion measures normally available. The second includes a variety of evidence of student competence that is typically available to graduate faculty but ordinarily not systematically assessed. The third category includes various types of developed competence which in theory could be assessed but would require a specific effort to do so. These three types are considered in some detail in the following sections.

Administrative Criteria

A number of criteria have been used for years in the assessment of graduate student performance. These criteria, which we have chosen to label "administrative criteria," include such indices of student competence as grades, performance on qualifying and/or comprehensive examinations, degree status (progress toward the degree, whether one eventually earns the degree), and dissertation quality.

Grades

When people speak of success in graduate school or "the criterion" of successful graduate student performance, more often than not they are referring to grades in one form or another. Along with the criterion of degree attainment, grades have been used more than any other criterion in studies of graduate school success or validity of the Graduate Record Examinations (Willingham, 1974).

As an indicator of student performance, grades have several positive qualities. First of all, they presumably represent the faculty's view of how the student is doing academically. Also, they are usually readily available for virtually all students, and therefore make a very convenient criterion. In fact, in a recent validity studies project carried out in cooperation with more than 30 graduate schools, Wilson (1978) reports that the first-year grade point average is the only criterion that is common to all institutions. Furthermore, in spite of an upsurge of interest in the use of nontraditional grading procedures at the graduate level in the late 60's (Benson, 1969), it is still the case that the form grades take--that is, the scale used to summarize grades--is still predominantly an A-F or five-point scale with clear meaning to most observers. As one graduate dean noted, grades serve as a shorthand communication that is presumably useful to the student, the teachers, college and university administrators, and prospective employers (Dale, 1969).

Grade-point averages do seem to represent a good composite of whatever kinds of academic performance are reflected in grades, since variation in student performance across a large number of courses can be accounted for fairly well by one general achievement factor (Boldt, 1970; French, 1951). Further evidence that it is reasonable to treat grades as representing a single, general kind of academic performance is available from studies done at the undergraduate level (e.g., Clark, 1964; Barritt, 1966). Thus, even though there is both empirical and anecdotal evidence that different teachers weight student qualities differently when assessing student performance--qualities, for example, such as student effort, amount of improvement during the term, clarity of expression, level of curiosity, etc.--it nevertheless appears that a large part of the information in grade averages can be explained by some unidimensional concept.

Another advantage often claimed for grades is their stability or consistency across terms. That is, students who earn high grades during the first term are more likely to earn higher grades during later terms. This is definitely true at the undergraduate level, though perhaps not so dramatically as many observers might think; although the similarity in academic performance between back-to-back academic terms is fairly high (with correlations between adjacent-terms grades often running in the .60′s and .70′s), grades over an extended period of time are much less stable (Juola, 1964; Humphreys, 1968). At the graduate level evidence regarding the stability of grades is more difficult to find. It is clear that there is less fluctation in grades at this point simply because almost all students receive A′s and B′s, but such consistency does not necessarily imply reliable measurement.

Difficulties with grades as a criterion in assessing student performance are numerous. First, it should be recognized that assigning grades among very able students, often working toward somewhat different individual objectives, is often very difficult under the best of circumstances. But the circumstances present additional difficulties. One technical difficulty is that the narrow range in grades assigned attenuates the magnitude of validity coefficients when grades are employed as the criterion in prediction studies. More importantly, the restricted range means that grade differences among students do not fully represent the range of differences in student accomplishment. This is especially true when some portion of the variance in grades can be explained on the basis of characteristics of the graders (e.g., faculty preferences for research papers that are experimental rather than descriptive) rather than achievement of the students. Grades at the graduate level thus may not provide meaningful descriptions of differential student performance, yet they are frequently used in determining the allocation of opportunities and rewards on the assumption that they report something specific and significant (Sparks, 1969).

A second shortcoming of grades is the obvious fact that grading standards can and do vary dramatically and sometimes arbitrarily across disciplines and within disciplines across different institutions (Bowers, 1967; Goldman and Slaughter, 1976; Juola, 1968). As a result, grades are practically useless as a criterion for multi-institutional comparative studies of student performance. Additionally, different grading standards means that special statistical techniques are necessary (Wilson, 1978) in order to combine data across institutions (within the same discipline) for validity studies, a strategy that is sometimes desirable owing to the small number of students within one department. Pooled data that does not make adjustment for such scale differences can sometimes result in an overall negative relationship between the predictor and the criterion, even when the "true" relationship as revealed in the various single-department (non-pooled) analyses is positive.

A third difficulty with grades as a criterion is that it is not always clear what grades mean. Different professors value different types of achievement. In spite of the finding cited earlier that course grades can be accounted for by a fairly general achievement factor (Boldt, 1970), it is at the same time true that grade assignment is sometimes unduly influenced by student characteristics that bear no clear relationship to academic performance, such as gregariousness (Singer, 1964) or gender (Caldwell & Hartnett, 1967). Furthermore, first-year grades in graduate school have been found to be only slightly related to eventual success in doctoral work in psychology (Hackman, Wiggins, and Bass, 1970), and it is likely that the basis for grading is quite different before and after students are accepted to formal candidacy.

It should be emphasized here that while these are reasons to be concerned about grades as a criterion of graduate student performance, these shortcomings should not be regarded as arguments for the abandonment of grades in graduate education. Perhaps such an argument can be made, but the evidence cited here hardly makes a good case for such a position, nor is it meant to. Grades serve several different purposes and their failings as a criterion may have little or nothing to do with their utility in other ways. This point was made frequently by graduate faculty members in our interviews. Virtually all graduate faculty members with whom we spoke recognized the limitations of grades as summarized on these pages, but few if any of them favored doing away with grades. Several argued, for example, that grades serve as a useful motivator for student performance, an argument that has some empirical support at the graduate level (Clark, 1969). Others pointed out that grades serve a useful feedback function by informing students of their progress, and still others took the position that as long as student evaluation was important and expected, grades probably provide about as useful a summary statement as anything else. In this respect, their views seemed to agree with Warren (1970), whose extensive review of the research on grading led him to conclude that grades probably do represent something useful, but it is not at all clear what it is.

Our conclusion, then, is that while grades serve several useful functions in graduate education, the one served least well is that of providing an understandable criterion of graduate student performance. Even here, however, it is important to remember that as one of several indicators they are relevant and useful. Their numerous limitations as a sole criterion should not detract from their value as one piece of the overall student evaluation puzzle.

Furthermore, it seems to us that there are several things that can be done at the department level that would make grades more valuable as an index of student performance. First, and perhaps most important, serious consideration should be given to the question of what purpose grades are expected to serve at the doctoral level. In his extensive review of undergraduate grading practices, Warren (1970) makes the observation that the purpose of grades has often been ignored. This criticism would seem to apply to the graduate level as well. Are grades intended primarily to motivate students? To provide students with useful information about their performance? To provide the department with a convenient number to use in making decisions about student eligibility for formal admission to degree candidacy, financial aid, and the like? To communicate valuable information to prospective employers when students leave the program?

Careful consideration of these various potential functions grades <u>can</u> serve, may well improve the purposes grades <u>do</u> serve. At the very least, it would encourage more graduate departments to reconsider the form grades should take, and whether one form can realistically serve various purposes equally well. Such an analysis would also make it necessary to rethink the relationship between grades and other procedures used to assess student performance. For example, if grades are meant to serve as summative evaluation statements about student progress in their courses, seminars, laboratories,etc., how do they differ from the role of the comprehensive examination?

Another step that graduate programs might take with respect to improving grading practices has to do with the function of grades as summary statements of student achievement and performance in basic coursework. The utility of grades for this purpose would be enhanced by urging faculty members to make a greater number of distinctions among students where such achievement distinctions occur and can be reliably assessed. The current tendency, as reported earlier in this section of the paper, is for the great majority of graduate faculty members to assign only A or B grades. When the student performance being assessed is extremely homoeneous, such a practice makes sense, of course. But when varied student achievement is the case, more varied and discriminating summary evaluations would make grades far more useful as criteria of performance. If there is great resistance to the use of C grades, discrimination in student performance can still be achieved by careful distinctions from A+ to B-, by the use of numerical rather than letter grading, or use of some other forms of grades.

These steps would not eliminate the problems of grades at the graduate level, but they would constitute a movement in the direction of shaping grades in the best interests of both graduate students and graduate departments.

Degree Attainment

Whether or not one earns the degree has frequently been used as a criterion for validating graduate school admissions criteria. In fact, degree attainment has been employed in validity studies as often as the grade-point average (Willingham, 1974).

There are several things to recommend degree attainment as a useful and important criterion of graduate student performance. For one thing, it is generally regarded as the single most important criterion of success by many, if not most, observers. Those who take this position argue that all other administrative criteria--grades, faculty ratings, or whatever--are simply poor proxies for what really counts; namely, did the student eventually earn his or her degree. Graduate students clearly regard it as the most important outcome of their graduate studies.

A careful analysis of degree attainment as a performance criterion requires an understanding of the factors related to student failure to complete the doctorate. Surprisingly, not much is known about the magnitude of attrition at the doctoral level, and what little is known mostly is the result of studies that were conducted some time ago. One of the most frequently cited studies of attrition in graduate school (Tucker, Gottlieb, & Pease, 1964) defined dropouts as all students who had not earned their degrees at the same institution within ten years. Using this definition, they found attrition rates averaging about 38 percent, a figure remarkably close to the 40 percent estimate given by graduate deans in a survey reported by Berelson just a few years earlier (Berelson, 1960). In a more recent study of attrition among doctoral Woodrow Wilson Fellows, Mooney (1967) found an attrition rate over 50 percent, a figure that comes very close to estimates made by Breneman (1977). However, both Mooney & Breneman defined a dropout as any student not completing the degree within eight years, compared to a ten year span for Tucker, et al. More important, however, Mooney found the doctoral-level attrition rate to vary dramatically by institution, discipline, and student gender. In reviewing these findings, a report of the Carnegie Commission on Higher Education concluded that "...from one-third to two-thirds of the population of prospective doctoral students will in fact earn their Ph.D. within an 8 to 10 year time span. The percentage of success will be highest (70 to 80 percent) for males studying physical science or psychology at a small distinguished graduate school and very low (5 to 15 percent) for women studying humanities and the arts at a large but less distinguished university" (Spurr, 1970, p. 128).

If these figures are reasonable estimates of the attrition rate picture in the late 1970s--and we remind the reader again that there has been no major examination of the question in more than 10 years-- then it is clear that doctoral level attrition is a matter deserving of careful scrutiny in its own right. As a criterion for the study of graduate student performance, however, degree attainment has certain limitations. For one thing, it requires a great deal of patience on the part of anyone collecting the information, for it is often many years before one can say with certainty whether or not a given student earned the degree or stopped trying. It is not uncommon, for example, for students in some fields to spend over 10 years working on an advanced degree, a matter which will be discussed in greater detail in the next section of this paper.

Beyond this logistical shortcoming, however, there is another difficulty with degree attainment as a criterion. Students drop out of graduate school for a host of reasons, many of which have little or nothing to do with competence or academic ability. Research indicates that graduate students frequently withdraw for reasons having to do with emotional problems (Halleck, 1976); poor relations with their faculty advisor (Heiss, 1970); family, health, or financial problems (Tucker, Gottlieb, & Pease, 1964); and so on. Worse yet, the real reason for withdrawing may never be learned. As Berelson (1960) points out, there is sometimes--how frequent we cannot say--a discrepancy between the real reason and the reason reported by those withdrawing.

The fact that so many different factors are involved in attrition means that the collection of useful information is more troublesome than one might have originally expected, and also that one needs to be very cautious in making inferences about the validity of selection measures when degree completion is used as a criterion. Knowing that many students drop out of doctoral programs because of lack of interest in the discipline (rather than, say, intellectual shortcomings) is useful in at least two respects. First, it underscores the inevitable fact that predictors that are primarily academic in nature--undergraduate grades and GRE scores, for example--will rarely be highly correlated with doctoral attainment when only those with the highest grades and test scores are admitted to doctoral programs in the first place. Second, such findings help us understand what other sorts of human qualities are required for successful performance in high-level academic activities, and thereby suggest a number of relevant human characteristics to assess as potentially useful additional predictors.

This general notion was brought up frequently in discussions with graduate deans and faculty members, who pointed out that, from their perspective, intellectual ability is seldom a very important factor in accounting for the difference between doctoral students who do and do not complete the degree. In their view, care is taken to insure that all admitted students are very able intellectually; thus, subsequent withdrawal is usually due more to personal qualities that are generally not tapped--or, at least, not assessed in any careful, formal way--when admitting students.

Two other observations about attrition were picked up in discussions with graduate faculty members and deans. First, it was almost always the case that faculty member's explanations for student attrition were reasons that were the "fault" of the students', not the department's. They rarely saw aspects of the program itself as having more than minor influence on a student's withdrawal decision, and chose, instead, to focus on personal qualities and characteristics of the student (e.g., lack of motivation, personal problems, etc.). Such an inclination conflicts sharply with evidence regarding the very important role of the department (or, in some cases, the university) in such student decisions (Heiss, 1970; Katz & Hartnett, 1976).

A second fact that became apparent in these discussions is that most graduate programs keep very inadequate records about the attrition question, an unfortunate state of affairs that was confirmed in a recent national survey (Clark, Hartnett, & Baird, 1976). The fact that departmental records are usually inadequate in this area is understandable, for the whole question of defining a doctoral-level dropout is not at all simple. As hinted earlier, often cases of dropping-out at the doctoral level are less matters of a definite, formal decision on the students' part, than a long-term <u>indecision</u> process that results in failure to re-enroll and which, after a time lapse of several years, is recognized as a de facto withdrawal without any kind of official (or sometimes even informal) communication of intent. In addition, the true reasons for student withdrawal are seldom known, sometimes because withdrawing students aren't sure themselves, but more often because they are reluctant to tell others the true reason, especially if the "others" in question happen to be in some official capacity with the institution. As

Berelson (1960) remarked, "What is critical frankness between doctoral candidates becomes in the dean's office lack of funds or personal change of plans" (p. 170). Obtaining such information from withdrawn students via anonymous questionnaire (ala Tucker, Gottlieb, & Pease, 1964) would seem to present less of a problem in this regard, though it is undoubtedly true that a certain amount of distortion of the real facts are present in data collected this way as well.

Still, it seems to us that most graduate programs could and should make more concerted efforts to routinely collect information from their doctoral students that would make it possible to periodically examine both the extent and major causes of attrition. As far as we can tell, the plain fact is that most doctoral programs--there are notable exceptions, to be sure--don't really pay very close attention to the simple question of how many admitted students eventually earn the degree, in spite of its obvious importance to the life of the department. In spite of the several shortcomings of degree completion as a criterion of doctoral student performance, it has numerous advantages and a compelling logical and intuitive appeal. A new national study of attrition would seem to be a useful step, though it might be wise first to develop improved models for persistence studies so that cumulative statistics would be more informative and better represent the basis for attrition.

Time to the Degree

The time-span between beginning doctoral study and completing the requirements for the degree has been a much-criticized aspect of advanced study in this country. The time-span problem--referred to frequently as the "drag-out problem"(Spurr, 1970) or "The Ph.D. Stretch-Out" (Wright, 1957)--has received considerable attention from educational researchers, both at the national level (Tucker, Gottlieb, & Pease, 1964; Wilson, 1965; National Academy of Sciences, 1967) and at various doctoral-granting universities such as Columbia (Rosenhaupt, 1958), Michigan (Bretsch, 1965; Heine, 1976), and Harvard (Doermann, 1968). These studies often report different results, largely because they vary with respect to the way time-to-the-degree was measured. The highest median figures are reported for the elapsed time between receiving the bachelor's and receiving the doctorate, the next highest figures are reported for the elapsed time between entering a graduate program and receiving the degree, and the lowest elapsed time data are reported when the actual enrolled time is used as the measure.

To complicate matters still further, time-to-the-degree data vary dramatically by discipline. For example, Wilson (1965) reported a median time span (between entering a doctoral program and receiving a degree) of 11.3 years in English compared to 6.4 years in psychology, and the National Research Council (1968) found a B.A.-to-Ph.D. time span of 11.6 years in Religion and Theology compared to 5.5 years in chemistry. Generally, the data seem to suggest that the time span differences are lowest in the sciences, followed by the social sciences, humanities, and professional fields (particularly education) in ascending order of median time-to-the-degree.

Time-to-the-degree data have been used occasionally as a criterion in studies of success in graduate school. Willingham (1974), for example, found more than a dozen studies in which time-to-the-degree was used as a criterion in prediction studies with GRE test scores. How long it takes one to earn the degree, like degree attainment, does have a certain rational appeal. The speed with which one accomplishes complex tasks has always commanded respect in academic circles, and it is probably reasonable to surmise that, within a given discipline, those who complete all degree requirements in three years are more able, on the average, than those who take six years.

The major drawback of time-to-the degree as a criterion, however, is that the reasons for taking longer are often ones over which the student has little or no control. It is true that some students don't earn the degree sooner because they can't, simply because they have difficulties meeting the requirements for the degree. In these cases, time-to-the-degree would appear to be a clear function of intellectual ability, willingness to work, "staying power," or other similar characteristics, and is thus a logical criterion. But in many other cases time-to-the-degree is a function of financial stringency (requiring the student to work at something other than completing the dissertation, for example), difficulties with dissertation committees (especially in the form of prolonged absences from the campus), and so on. Berelson (1960) even suggests that some students actually aren't allowed to finish sooner because "...they are needed as teaching assistants for the department or research assistants for the professor" (p. 162).

For this reason, then, time-to-the-degree would appear to be weak as a single criterion for assessing graduate student performance, a conclusion with which most graduate deans and faculty members seem to agree. If degree duration were more a function of student abilities and motivation, we would feel more favorable toward it, but on the basis of both the research evidence and opinions of graduate faculty members, that does not appear to be the case. On the other hand, it is clearly a relevant datum to include as one of a number of relevant criteria. In addition, time lapse information can be informative and useful as an indicator of program functioning, particularly if the information is analyzed over time. For example, if students are taking considerably longer to complete their degrees than students in the same discipline at other institutions, or longer now than ten years ago in the same department, this information could turn out to be a useful red flag for purposes of program evaluation and improvement.

Comprehensive Examinations

The nature and form of comprehensive examinations varies considerably, both across disciplines and across institutions within a discipline. More often than not, however, the term "comprehensive examinations" applies to an examination or set of examinations--usually written, occasionally oral--that follows the student's completion of formal course work at the graduate level and is used to determine the student's mastery of research in the field and eligibility for formal degree candidacy in the department.

In some institutions these are referred to as "qualifying examinations." With some exceptions, the only difference is the name, not the timing or basic purpose of the test. Therefore we use the terms "comprehensive examinations" and "qualifying examination" interchangeably. It should also be noted that the term "comprehensive examination" is often used in a more general sense to refer to the overall evaluation of a student. In that event, other types of information are included in a broader, and necessarily more subjective assessment of the students' status. Our use of the term here focuses on the more narrow interpretation, i.e., a specific examination.

One of the most commonly criticized weaknesses of comprehensive examinations--at least as practiced at most institutions--is the frequent departmental uncertainty and lack of specificity about the purpose of comprehensive examinations and, consequently, their basic form and content. As one critic observed, "...graduate departments in many cases have never defined for themselves, much less for the students, what ground the examination should cover and how to go about preparing for it" (Carmichael, 1961, p. 149). Apparently, this observation, made nearly 20 years ago, is still an accurate description of the status of doctoral-level comprehensive examinations (Mayhew & Ford, 1975).

Our discussions with graduate faculty members suggest that some departments do not take the comprehensive examination very seriously, and are apparently not very concerned about taking steps that would make them more reliable and meaningful measures of student attainments. Several surveys, in fact, attest to the general lack of concern among graduate faculty members about the comprehensive examination (Heiss, 1970; Berelson, 1960). Thus, it is not uncommon for departments to fail to clarify--either for students or themselves--what the purposes of the examinations are and what general content they will cover. Pre-examination syllabi or bibliographies of major reading material still seem to be fairly rare. In one study at ten major universities, over 50 percent of the graduate students pointed out that they received no briefing from their advisers about the likely content of the exams (Heiss, 1970). In addition, the content of the tests often varies dramatically from year-to-year, complicating the matter still further. The effect of this practice on students in the form of increased anxiety can only be guessed at, though available anecdotal evidence (Heiss, 1970; Katz & Hartnett, 1976) suggests that it is neither a minor nor infrequent problem.

In addition to difficulties with the purposes and content of comprehensive examinations, evidently few graduate departments have given serious attention to the question of how to grade such exams, which are almost always in essay or expository form. As a result, it may well be the case--we have no evidence for this assertion--that evaluations of student comprehensive examination performance are often not very reliable.

Most graduate faculty members seem to be aware of the above-mentioned shortcomings with current comprehensive examination practices, and some concede that their practices in this regard have not been very careful. At the same time, there appears to be a reluctance to consider more systematic procedures for the assessment of student academic attainment, a reluctance that was also expressed in a 1971 survey of opinion about the feasibility of a common criterion (Carlson, Evans, and Keykendall, 1973). To some extent, lack of concern about the soundness of assessment stems from the view that doctoral program admission procedures yield such highly able students that there is simply no need to be concerned about careful and time-consuming subsequent evaluations. To some extent too, it is probably a reflection of an unspoken aversion to "standardizing" the graduate student assessment process. And finally, there is the position that the graduate student learning experiences are so diverse that it is simply not possible to spell out in any detail the kinds of competencies that students should possess. Therefore, the suggestion that graduate faculties should specify the competencies they expect of students and construct examinations to test whether those competencies have been achieved is seldom given serious consideration, and examples of the success of such practices in professional fields (e.g., Rimoldi, 1963; McGuire & Babbott, 1967) are often regarded as irrelevant.

In view of the apprenticeship nature of advanced graduate studies in most fields, it is not difficult to understand and appreciate faculty member's concerns about the undesirability of examinations which would pressure departments in the direction of requiring uniform academic skills and competencies. Nevertheless, it seems to us that this concern is somewhat misdirected. After all, comprehensive examinations precede the dissertation and other important training that is clearly individualized. The purpose of the exam, as employed at most institutions at least, is to assess the student's comprehension of the basic facts, methods and theories of the field. In many departments, comprehensive examination performance constitutes one of the major criteria used by the faculty to make judgments about student qualifications and potential. Students who withdraw from (or are "encouraged out" of) doctoral programs for academic reasons often do so at this point. Obviously, then, it is important that graduate departments pay close attention to ways of insuring that the content of such exams matches the objectives held for them, and that the procedures for evaluating student performance result in fair and accurate assessments and also provide helpful feedback to students.

Dissertation Quality

The dissertation is the culmination of the doctoral program and its quality receives a great deal of attention by graduate faculty. Evaluation of dissertation quality and the decision to award the degree are necessarily somewhat subjective. It is surprising, however, that the dissertation has not received more attention in validity studies and formal evaluations of graduate programs. The dissertation uniformly stands as the primary piece of evidence that a student can conduct sound scholarly and research endeavors, and in spite of Barzun's (1968) claim that the dissertation is beyond many students' strength financially, socially, and emotionally, the evidence is clear that the dissertation is highly valued by both students and faculty (Berelson, 1960; Porter & Wolfle, 1975). As further evidence of their general esteem, many disciplines conduct annual competitions to identify particularly outstanding dissertations. Nevertheless, with the exception of several non-resident doctoral programs (e.g., Medsker & Wattenbarger, 1976; Meeth & Wattenbarger, 1974), there has been scant attention to the dissertation as a useful indicator of comparative doctoral student performance.

There are several positive aspects of the dissertation as a criterion. First as already indicated, it is regarded as the central test of ability to carry out scholarly activities. Furthermore, it has a "real-life" appeal that is undeniable, for in its properly monitored form, it tests the extent to which students can conceive and carry out activities that are expected to occupy a substantial part of their professional lives. Completing the dissertation requires much more than academic ability. Though cognitive skills are clearly necessary, without such qualities as patience, perseverance, and the ability to overcome the fear of putting one's words on paper, dissertations would never get finished. We know of no good evidence dealing with the topic, but it is our impression that very able and bright graduate students who, for whatever combination of reasons, were unable to complete their dissertations--the well-known "ABD's" that are familiar around many colleges, universities, and research centers--are often the same individuals who, as academic staff members, are less likely to produce the necessary project reports, journal articles, and other products that characterize work in academic settings. For these reasons, completion of the dissertation does have considerable merit as a performance criterion.

In this respect, however, it should be pointed out that dissertation completion is virtually the same as degree completion, a criterion discussed in some detail earlier in this paper. The criterion being considered here is not degree completion, but degree quality. It is our feeling that numerous practical difficulties would be encountered in any attempt to employ dissertation quality as a criterion of research competence. For one thing, it would require the use of carefully selected objective panels of readers in the discipline, each of whom would have to read numerous dissertations and make ratings on standard, carefully-constructed dimensions of quality. Steps would obviously need to be taken to insure author anonymity, probably even mask institutional affiliation. Furthermore, there is the problem of lack of agreement among different readers (just as there is now, for example, for different reviewers of manuscripts submitted to professional journals), and this difficulty would have to be resolved. Thus, any such undertaking would be fairly expensive and time-consuming. Though such barriers might well be overcome in a carefully designed research project, it seems unlikely that ratings of dissertations in this manner could ever be expected to become routine practice.

In addition, it is sometimes difficult to know just what portion of the dissertation represents the work of the doctoral student and what portion is the work of the student's major professor. The final writing of the thesis, to be sure, can almost always be assumed to be the work of the student. But what about the major conceptual orientation or hypotheses of the inquiry, the basic research design or strategy, or methods chosen to analyze the data? To some extent, these considerations are always influenced by a student's dissertation chairperson and committee members. Berelson (1960), for example, reports that graduate students select their own dissertation topics very rarely--less than 10 percent of the time, in fact, in the humanities and sciences. The problem is that, even within a department, students are influenced unevenly, and therefore the extent to which the dissertation serves as a true measure of the student's research competence is not always clear.

By and large, the faculty members with whom we spoke shared these concerns about the utility of ratings of dissertation quality as a means of assessing performance. Though it would be useful to examine it as a dichotomous variable (completed vs. did not complete) in the same way that degree completion would be examined, attempts to obtain objective, unbiased ratings of dissertation quality would probably be too difficult and expensive.

This is not to say, of course, that examination of the dissertation as part of the educational process is unimportant. Dissertation quality should be a matter of definite concern to members of any academic department, and some sort of routine procedure for monitoring the overall quality of dissertations within a program would seem to be essential. The previously-mentioned reservations make it difficult to make individual distinctions, but some review procedure should be established to insure that a high standard is being maintained for the entire program.

One final observation is in order before we conclude our review of administrative criteria. Each of these criteria is "formal," in the sense that the evaluations tend to occur at prescribed dates (e.g., comprehensive examinations) or over a prescribed period of time (e.g., a course grade), and some summary result of the evaluation is then transmitted to the student so that it becomes part of both the student′s and the department′s official record. It needs to be recognized that a good deal of the evaluation process in graduate education--just how much, we cannot say--operates in a more informal fashion and its results never become part of any formal record. Several faculty members pointed out to us, for example, that faculty members tend to form opinions or make judgments about students after contacts with them over a long period of time in a variety of settings. Then, on the basis of these gradually-formed opinions, they give less support and encouragement to the less able students in the form of such things as: personal communications and contacts, invitations to join in collaborative research efforts, opportunities for teaching and research assistantships, unwillingness to serve on dissertation committees, discouragement during work on the dissertation, and the like. The course grades for these students may be acceptable (presumably because some faculty members are reluctant to assign poor grades due to the fact that their assessments will not be anonymous), and their performance on the comprehensive examination may have been acceptable (after several attempts), but, by means of these other more subtle mechanisms, such students are gradually "cooled out" of graduate study. To the extent such informal assessments actually occur and are communicated to students, it suggests that the administrative criteria reviewed here provide an incomplete picture of the way graduate student performance is evaluated.

Evidence of Professional Accomplishment

A substantial body of research literature has developed in recent years that deals with student accomplishments. Basically, this research indicates that self-reported accomplishments at one educational level (secondary school, for example) tend to predict similar accomplishments at a later educational level (e.g., college). Perhaps the best evidence comes from the National Merit Scholarship Corporation, which reported a series of studies in the 1960s clearly indicating that the best predictor of a specific nonacademic accomplishment in college (e.g., composing or arranging music which was publicly performed, getting elected to one or more student offices) was accomplishment in that same (or a very similar) area in secondary school, as measured by a simple student self-report from a checklist (Holland & Nichols, 1964; Nichols & Holland, 1963). Even more striking was the finding that such specific accomplishments are not accurately predicted by such standard academic indices as grades or verbal aptitude test scores (Baird & Richards, 1968; Richards, Holland & Lutz, 1967; Wing & Wallach, 1971).

The Graduate Record Examinations Board has been interested in student academic and nonacademic accomplishments, and has supported one project designed to develop an experimental inventory of documented undergraduate accomplishments for use in graduate school admissions (Baird, 1976b). However, the content of such a form has naturally been concerned with college-level accomplishments and activities. Comparable self-report forms could of course be developed for use in documenting graduate-level accomplishments that might reasonably be expected to occur during the student's graduate career and would be relevant to scholarly or professional performance.

One step in this direction was taken recently when two ETS researchers designed a graduate student accomplishments questionnaire to gather criterion information in a study of the validity of experimental Tests of Scientific Thinking (Ward & Frederiksen, 1977). Among other things, students were asked to indicate whether they had attended meetings of a scholarly or professional society, subscribed to two or more scholarly or professional journals, authored or coauthored scientific papers submitted to or published by professional journals, and so on. A major finding of this research was that such professional accomplishments were more accurately predicted by the experimental Tests of Scientific Thinking than by GRE test scores.

Such indices of professional behavior have considerable merit as criteria because they reflect important long-term objectives. If such measures are to be seriously considered as a graduate student performance criterion, routine procedures for the collection of such information would seem to be essential. Currently, such student accomplishment information is rarely kept in any systematic way in departmental files. In addition, for such information to be used in research involving multi-institutional comparisons of student performance, it would be necessary to collect such data on standard forms, using common definitions of accomplishment and data formatting procedures. The net result might be a sort of vita for students that would be standardized with regard to terms and accomplishments deserving of recognition, but at the same time flexible with regard to the range of achievements that could be reported. Obviously, such a form would have to differ by academic discipline. It could be either a completely self-report document, a record kept by someone in the department office (perhaps being checked and verified by the student), or some combination of these.

Note that the purpose of such information would be to facilitate research on questions having to do with student performance. Such data would not be intended for use, in the early stages at least, for evaluating individual students. The routine collection of such information from graduate students by means of some systematic procedure(s) would make several avenues of research possible at the departmental level. First, taking the lead from previous research at the undergraduate level (e.g., Richards, Holland, & Lutz, 1967; Wing & Wallach, 1971), departments could conduct local studies of the relationship between admissions variables, routinely available academic criteria, and subsequent indices of student professional accomplishment. By doing so, they might learn answers to such questions as whether test scores and undergraduate grades are useful predictors of which students will author or coauthor papers or whether any of the selection variables forecast which students make outstanding teaching assistants. Second, information about student accomplishments in graduate school would be important to examine periodically in its own right as a way of assessing the professional development and academic socialization of students within the department. This might be particularly useful if similar information were available for comparison purposes from a variety of other departments within the same discipline. In the latter instance it would be possible to determine whether students in one department seem to be presenting papers at regional meetings as frequently as students from another institution, whether there is any evidence that faculty members in the department are not drawing graduate students into meaningful (and visible) collaboration efforts to the same extent that this is happening in other programs, and so on.

As with many of the other possible criteria of graduate student performance, however, it is clear that there would be several significant limitations with student professional accomplishments. One is that such accomplishments may be partly a matter of simple luck. In our discussions with graduate faculty members, many of them pointed out that some graduate students publish journal articles as joint authors or coauthor papers presented at professional meetings because they happen to be fortunate enough to be associated with a major professor who is nurturant and supportive in these regards, whereas other students are perhaps equally competent but do not receive the same encouragement or assistance. Similarly, while one student′s work on a project may result in coauthorship of a journal article with one professor, an even more substantial contribution from another student may not even earn an "I am indebted" footnote from one more selfish. To the extent that such differences are commonplace, these kinds of student behaviors are misleading as indices of individual student accomplishment.

A second difficulty with the professional accomplishments criterion is essentially a psychometric limitation--namely, that the distribution of such accomplishments will be extremely narrow and skewed. At least this is true at the undergraduate level (Richards, Holland, and Lutz, 1967), and is about surely the case in graduate school as well. Several faculty members we spoke with suggested that the "typical" graduate student will have published no papers at all (nor presented a paper at a professional meeting) and that it will be extremely unusual to find students who have done so more than once or twice. On the basis of what is known about graduate education such an assessment would seem to be fairly accurate, at least in the social sciences. This does not affect the logic of using accomplishments as a criterion, of course, but it does reduce their likely utility.

In spite of these shortcomings, it still seems to us that professional accomplishments offer a criterion possibility worthy of further serious consideration. As is true of any evaluation, one should not make too much of a particular incident, but, rather, assess the overall pattern that a student′s graduate career suggests. As we mentioned earlier in the paper, there is a fairly widespread feeling among graduate faculty members interviewed that the "better" graduate students aren′t necessarily brighter or academically more able than their peers (all of whom, we were frequently reminded, were carefully selected on such variables), but were students whose motivation and drive were outstanding as evidenced by various types of incipient professional behavior. To the extent that this is so, a professional accomplishments criterion is both relevant and promising, in spite of the several shortcomings discussed earlier. In fact, the reluctance among some faculty members to endorse an accomplishments criterion, in view of their feelings about the importance of professional socialization, seems to be an example of circular reasoning. This is especially the case since it seems clear that most faculty members, in one way or another, do seem to pay attention to evidence of student professional socialization. What it may come down to is that their concern is less with the criterion than with the idea of developing "objective" or "standard" means of measuring or recording the information. Many faculty members, especially in the arts and humanities, seem to have a rather pervasive suspicion of almost any attempt to quantify or routinize information that they regard as important.

We would hope that more graduate programs would pay closer attention to evidence of student professional accomplishments. It may turn out that such accomplishments are simply too rare to serve as a useful index of student performance or that, for other reasons, it is not a feasible criterion. But it would not take much time or effort to explore the possibility, and it just may result in information that will prove to be useful to departments in better understanding both the students′ performance and their own.

Specially Constructed Criteria

In addition to the standard administrative criteria employed in graduate student evaluation and student self-reported documented accomplishments, there is a third category of criterion information that needs to be considered. These are specially constructed measures of various critical competencies regarded as important outcomes of advanced training, but outcomes or competencies which are rarely assessed in any systematic way by most graduate programs. In this section we consider four different types of constructed criteria: (1) assessments depending on gaming, simulation, and role playing; (2) new tests of academic achievement or problem-solving abilities; (3) rating scales; and (4) performance work samples.

Gaming, Simulation, Role-Playing

Certain gaming, simulation, and role-playing exercises, similar to those that have been utilized in management and military exercises, would seem to be appropriate to consider for use in educational assessment. In fact, in certain situations, especially medical education, carefully-developed simulation exercises have been found to be fairly reliable (Lewy & McGuire, 1966) and valid for certain purposes (McGuire & Babbott, 1967). As a general rule, however, these simulation exercises more closely resemble work performance measures which we shall discuss in a later section. Most gaming and simulation activities, on the other hand, have several serious disadvantages when considered as candidates for use in graduate student performance assessment. First, they have been generally conceived and implemented as instructional rather than evaluative procedures. As a result, they tend to have been developed with a far greater emphasis on demonstration variables (e.g., clear rules, realistic situations, simple physical equipment, easy observation) than on such assessment qualities as reliability, validity, and the like (Parry, 1971; Tansey & Unwin, 1969; Wildberger, 1974).

Second, when gaming, role-playing, and simulation exercises have been used for evaluation purposes, they have either been too unreliable to be acceptable for assessment purposes, or, in attempting to resolve some of these measurement concerns, they have usually had to resort to procedures (e.g., large numbers of specially-trained judges) that make their use both cumbersome and expensive (Inbar & Stoll, 1972).

Third, there is a definite clinical nature to such procedures which makes them inappropriate for use in most disciplines.

Currently, gaming and role-playing are rarely used in assessing graduate student performance. Due to the above considerations they are not likely to be widely adopted or used in the future. In certain situations or areas of study they might prove to be occasionally beneficial, but they are unlikely to be very generally useful to the graduate education community. Simulation activities that resemble work samples will be considered in a later section of the paper.

New Tests of Academic Achievement or Problem-Solving Abilities

Achievement tests. Those interested in the assessment of individual student or program performance at the undergraduate level have had a definite advantage over those with similar interests at the graduate level. Standardized multiple-choice measures of undergraduate student competence and abilities have been available for years, and often such measures are administered to college seniors as a way of assessing either student development during college or the general level of student competence. Such instruments have made it possible to measure student achievement both in a specific discipline (e.g., European History) as well as general education areas (e.g., humanities, science). One result is that we have witnessed the growth of a considerable body of research literature dealing with the impact of the undergraduate experience on students' cognitive development.

No such tests have ever been developed for broad use at the graduate level, and for good reason. The special apprenticeship nature of graduate training in most disciplines would make it extremely difficult to develop achievement measures that would have content validity as program outcome measures for most students. Much of the activities we think of as encompassing a Ph.D. program are non-course activities, such as independent study for the language requirements, work on the dissertation, and the kinds of experience, knowledge and skills picked up as a teaching or research assistant in the department. There is tremendous variation in the specific nature of these experiences even within one department, and it would therefore seem unreasonable to think of developing any kind of objective, standardized achievement test that would attempt to assess the outcomes of these diverse experiences.

However, such achievement tests do seem to have potential for use in conjunction with the comprehensive examination, though there would be serious problems here as well, because of variations in objectives and the core program from one department to another. "Local" examinations can reflect these differences, but other problems emerge; we refer the reader again to the previous discussion of comprehensive examinations.

Problem-Solving Abilities. The possibility of developing measures of certain problem-solving skills that would be appropriate for use at the doctoral level is an intriguing one. Educators have been convinced for years that one of the most important outcomes of education is a developed (trained) ability to think critically and analytically. Beyond the acquisition of knowledge, it has often been claimed that what one should hope to garner from education (especially, it is often argued, advanced or "higher" education) is an ability to reason carefully, to recognize valid and invalid inferences and assumptions, to approach problems with an attitude of inquiry that demands evidence in support of assertions, to identify logical contradictions in arguments, and so on. Unfortunately, efforts to assess the postulated construct or constructs have not been particularly successful.

One of the earliest and best-known efforts to assess non-factual intellective outcomes at the undergraduate level was the Cooperative Study of Evaluation in General Education of the American Council on Education, carried out in the late 1940s and early 1950s. A central idea that shaped the work of the ACE study was the concept of critical thinking. Thus, the final report of the project (Dressel & Mayhew, 1954), placed understandable emphasis on A Test of Critical Thinking. Unfortunately, the Test of Critical Thinking did not prove to be the promising breakthrough that many were hoping for, simply because, in spite of efforts to emphasize reasoning and problem-solving skills, the statistical evidence suggested that it was just another measure of verbal intelligence. As Dyer (1956) remarked "...verbal reasoning, such as one would find in an ordinary reading comprehension test, accounts for most of the common variance..."

A more recent attempt to assess critical thinking led to the development of the Critical Thinking Appraisal test (Watson & Glaser, 1964). The underlying construct in this measure is very similar (if not identical) to that described by Dressel, Mayhew, and their colleagues. Once again the research hints strongly that the CTA is largely a measure of verbal ability though there is some evidence that the CTA is also measuring something else.

Shortly after the ACE Cooperative Study of General Education was completed, a group of measurement specialists who were interested in measuring educational outcomes produced a taxonomy of educational objectives designed to help educators clarify their goals and improve their assessment of educational outcomes (Bloom, Engelhart, Furst, Hill, and Krathwohl, 1956). This team of measurement specialists devised a hierarchical taxonomy of objectives in which the objectives of one class are likely to make use of and be built on the behaviors found in preceding (lower) classes. There were six major classes of objectives. In order ranking from simple to complex they are: knowledge, comprehension, application, analysis, synthesis, and evaluation. Even without the specific definitions of these taxonomic categories, the reader can see that some of them--particularly the last three--look very much like some of the words used in the previously described efforts to measure critical thinking.

There was a flurry of research activity on the cognitive taxonomy in the late 50s and early 1960s, most of it aimed at trying to understand the meaning of the various categories. But in the mid-1960s less attention was given to the taxonomy or the concept of critical thinking, reflecting an accompanying decline in the level of interest in "general education" at the undergraduate level. As interest in general education waned, it is hardly surprising that efforts to assess the outcomes of general education became less frequent. This development, in conjunction with the problems of defining what was meant by critical thinking, logical reasoning, and so on, together help explain why so little advancement has been made during the past 20 years in understanding and assessing these elusive constructs.

In sum, critical thinking and similar generic competencies appear to be developed over long periods of time and are not ordinarily amenable to improvement through a graduate program unless one is focusing on a fairly specific type of problem-solving in a particular field. Thus, these types of competencies are usually thought of as aptitudes and treated as predictors rather than outcomes. The Analytic module of the GRE-Aptitude test is one recently-developed example.

Rating Scales

Global faculty ratings of graduate student performance have been used as a criterion measure in a fair number of validity studies, though they have not been employed in this way nearly so often as grades or degree attainment (Willingham, 1974). It would appear that ratings are an acceptable criterion measure, at least in many fields of graduate education (Carlson, Evans, & Kuykendall, 1973).

One advantage of ratings is that they are relatively easy to obtain, thus providing a fairly convenient criterion. Unfortunately, however, ratings still suffer from several serious shortcomings. Perhaps the most troublesome problem with ratings, at least as a criterion of graduate student performance, is simply that many members of the faculty will not be sufficiently familiar with the student′s work to be able to make an informed rating. This was evident in research conducted in graduate business schools (Hilton, Kendall, & Sprecher, 1970), and would seem likely to be characteristic of other graduate programs as well. In fact, when asked about this aspect of ratings, many graduate faculty members acknowledged that they often had only a casual familiarity with students other than their "own." To a large extent, of course, this depends on the size of the department.

In addition, ratings have often been beset with problems of leniency and range restriction (Reilly, 1974a). And though efforts to improve ratings through critical incident techniques did distinguish a small number of separate factors comprising graduate student performance (e.g., independence and initiative, conscientiousness, critical facility, etc.) in chemistry, English, and psychology (Reilly, 1974b), subsequent research revealed that scales developed to obtain ratings of these separate factors were highly intercorrelated and had only minimal reliability (Carlson, Reilly, Mahoney, & Casserly, 1976). The high intercorrelations were confirmed in research on undergraduate students, where it was found that faculty ratings of students are heavily dominated by an academic performance factor, as defined by grades (Davis, 1965).

Perhaps the most effective ratings scales are those which define the extremes of the behavior being observed and, if possible, also provide descriptions of intermediate points along the continuum. Such "behaviorally anchored" rating scales hold promise, but the utility of such measures depends heavily on the experience of the raters and the thoroughness with which they have been trained. Even with careful training, however, a "halo" effect--that is, the tendency for an observer′s general impression to influence his ratings of specific behaviors--and other forms of contamination are frequently difficult to eliminate when rating scales are used (Brogden & Taylor, 1950; Glaser & Klaus, 1962). Davis′ (1965) finding that faculty ratings of various traits of undergraduate students are all highly correlated with student grades is again relevant in this regard.

In certain respects, ratings have always been a fairly important aspect of student evaluation in graduate education and are likely to remain so. Grades, for example, are a form of ratings in one academic course (see discussion of the shortcomings of grades as a performance criterion earlier in the paper), and letters of recommendation another. Letters of recommendation, however, are almost always written by someone chosen by the student and therefore, presumably, by someone very familiar with the student′s work and abilities. Some departments apparently employ global faculty ratings in the process of making certain internal decisions (e.g., about student assistantships or certain field work experiences), but these ratings are rarely done in a very formal way involving the ratings of the entire faculty within the department. Given the problems of the lack of faculty contact with some students, rater unreliability, and halo effect, it seems unlikely that global faculty ratings will ever become an important or widely used criterion of graduate student performance.

One final aspect of ratings deserves to be mentioned. For research purposes, peer (fellow-student) ratings should not be overlooked, for they have been found to be promising predictors of subsequent performance, both in and outside of education. The usefulness of peer ratings for predicting success in the military was demonstrated many years ago (e.g., Bryant, 1956; Tupes, 1957 and 1959); more recently, their potential in educational settings was suggested when it was found that peer ratings of non-intellective traits were superior to both academic aptitude and self-report measures in the prediction of first-year performance in college (Smith, 1967).

At the graduate level the research on the utility of peer ratings has been infrequent but encouraging. Kelly and Fiske (1951) found peer ratings of clinical psychology trainees to be only slightly less accurate in predicting later success than ratings by trained psychologists, Eisenberg (1965) found peer ratings to be highly correlated with performance on comprehensive examinations in one doctoral program, and Wiggins and Blackburn (1969) found peer ratings to be better predictors of first-year performance in psychology at one institution than a host of other more traditional predictors.

These few studies identify peer ratings as a variable worth paying closer attention to as a potentially useful criterion. However, their reliability, construct validity, variation across disciplines, freedom from halo and other characteristics need to be more carefully examined. In addition, the problems involved in any administrative use of peer ratings would appear to be numerous. About all that can be said at this time is that peer ratings should be considered in research efforts to identify potentially useful ways of assessing attainments in graduate studies.

Performance Work Samples

In an address presented at the 1977 ETS Invitational Conference on Testing Problems, Frederiksen (1977) related the following situation. During the Second World War, it was found that the best tests for predicting grades in Gunners Mate school were verbal and reading comprehension tests, a somewhat surprising finding in view of what gunners mates were expected to do. Closer examination revealed that the primary reason was that grades were assigned on the basis of student performance on multiple-choice tests, whose content, in turn, was based on lectures and manuals. After tests were developed that actually required students to adjust and repair guns--tasks that were in line with the job for which the students were being trained--it was found that verbal and reading comprehension tests were no longer accurate predictors of the criterion performance.

This story has clear implications for the assessment of criterion performance in just about any setting, including graduate school; if we are interested in developing a measure of the effectiveness of a training program we need to first understand what kinds of skills and behaviors to expect from those who have experienced the training, and one way to accomplish this is to analyze the sorts of things graduates of these training programs will be doing in their subsequent careers and occupations.

One may understandably hesitate at the suggestion that we need a clearer understanding of the specific behaviors and activities that will be expected of the graduates of most graduate programs. These people, after all, will be employed in positions requiring very complex behaviors and skills, ones not easily defined nor simply described. It is one thing to give a precise description of the specific job activities of a lathe operator, quite another to so easily say what a college professor does. But detailed, minute descriptions are really not necessary. As Cronbach (1970) has argued, what is needed is not a test that will sample the criterion task exactly, "...but the general type of intellectual or motor performance required by the criterion task" (p. 199).

We do know that the primary purpose of the great majority of doctoral programs in this country is to prepare scholars and researchers (Clark, Hartnett, & Baird, 1976). Purposes such as the preparation of future teachers or practitioners are also acknowledged, but are not regarded as being nearly so important. In considering ways to develop additional systematic assessment methods, it is quite reasonable, then, to focus on the student's ability to carry out research and recognize the merits and deficiencies in the research reported by others. One possible way to assess the former is by closer, more objective evaluations of dissertations, a suggestion discussed in another section of this paper. An alternative way is by means of a specially constructed measure for each discipline that would directly assess important aspects of student research performance in a standardized task.

Research on the development of measures appropriate for use as criterion measures in advanced training programs is not a new area of inquiry. Previous research funded by the GRE Board, in fact, has resulted in the development of The Tests of Scientific Thinking, which are free-response job-sample tests which simulate tasks that might be encountered by a behavioral scientist (Ward & Frederiksen, 1977). Research with the experimental Tests of Scientific Thinking has indicated that the various TST sub-tests are not highly correlated with GRE scores and were more highly correlated than GRE tests with student self-reported professional accomplishments. These data suggest that relevant criterion measures for graduate student performance can be developed that are not simply extensions of traditional verbal skills measures.

It was this background that led the authors to conclude that work-sample criterion measures had great promise, and to the idea of possibly developing such an experimental measure. This would not be greatly different from tasks that graduate students encounter on examinations or in the course of their studies, but it was our belief that such graduate student work sample tests could be developed which could be standard tests suitable for use across departments within the same academic discipline. Such measures, we reasoned, would seem to have several advantages over routinely available criteria, in that they could:

--differentiate the various objectives or purposes of training (e.g., clinical vs. experimental psychology);
--reflect important types of competence or skill (thus serving as an appropriate "target" for a specialized predictor);
--provide comparative data; that is, data about student performance in various departments within the same academic discipline;
--allow better control of measurement characteristics (especially reliability and comparability).

Furthermore, such measures might well be attractive to many graduate programs, since they could be used not only as a means for improved student assessment, but also for purposes of program evaluation, in the design of curricula or training programs, in the validation of selection measures (and the improvement of selection procedures), in instruction, and even, possibly, for use in student guidance.

Our enthusiasm for the idea of a work-sample criterion measure, however, was not shared uniformly by graduate faculty members with whom we spoke about the possibility. Some felt this approach to assessment deserves more emphasis, but several other reservations were expressed.

Some argued that it would be difficult, if not impossible, to design a work-sample measure that would be appropriate to all Ph.D. candidates in a program or even branch of a discipline. Emphasizing the apprenticeship nature of the graduate experience, these faculty members argued that, in certain disciplines at least, there is not very much substance in common among the various subspecialties.

Also, such a standardized criterion measure might have the unfortunate effect of pressuring departments toward greater uniformity in their curricula. Given current student assessment procedures, within-department diversity is permitted to thrive, with less popular subspecialties often going their own way, eschewing pressures to adopt "the latest methodologies." In effect, this was an expression of concern about the extent to which a standardized criterion measure would gradually become the definer or undue influencer of the nature of graduate school curricula.

A third objection raised against the work-sample idea had to do with the problem of fit between the specific nature of the work-sample task and the sort of subsequent activities graduate students would be engaged in. A number of faculty members pointed out that, in view of the evidence regarding the relatively small number of Ph.D.s who in fact spend a substantial portion of their professional lives conducting research, a work-sample that assessed only how well students carry out research and recognize deficiencies in the research work of others would be somewhat inappropriate. Furthermore, it would be extremely difficult to know what sort of work-sample task would be appropriate and meaningful for most students in view of the great diversity of careers followed by recipients of the doctorate.

While there is merit in these points, we continue to feel that a work-sample approach has much to recommend it as one possible criterion of graduate student performance. It seems clear that the shortcomings many graduate faculty members perceive in such a criterion argue effectively against widespread use for administrative purposes, such as evaluating student performance across programs. One can still argue, however, that individual departments could use more work sample assessment to good effect, and that some research to further such use might be beneficial.

Conclusions and Recommendations

We have attempted a general analysis of the strengths and weaknesses of a fairly large number of criteria that have been or might be used to evaluate graduate (especially doctoral) student performance. Our original intention was to examine these criteria primarily in terms of their adequacy as dependent variables for the validation of graduate school selection procedures, but it soon became apparent that each separate criterion could be fully understood only after consideration of the more general process of graduate student evaluation. As a result, this report deals not only with the "criterion problem" in the traditional measurement sense, but also, to some extent provides an overview of student performance evaluation in graduate education.

The qualifying phrase "to some extent" is essential. We have not conducted a survey of doctoral program practices. Instead, our "method" was to read all that we could find that dealt with specific performance assessment practices and the more general question of student evaluation in graduate education, and to talk to a selected sample of informed and knowledgeable administrators and faculty members.

General Observations

Given the nature of our inquiry, we have no empirical evidence to present. Taken together, however, the literature and discussions have led us to numerous observations and general impressions about the practices that are being used to evaluate student performance in graduate education. Several of the more important observations are that:

--Very little research literature is available about how graduate student academic performance is assessed. Several general analyses of graduate education have dealt briefly with the topic (often in quite critical terms), but very little serious, thoughtful examination has been made of what does (and/or should) constitute successful student performance. Such a situation, we suggest, is related to and symptomatic of another general problem, which is that:

--Many graduate faculty members place low priority on systematic or formal efforts to evaluate graduate student progress or achievement outside the framework of existing practices. This is not a universal attitude by any means, but it is shared by many if not most faculty members and the reasons for it (some of which are discussed earlier in this paper) are extremely complex. One of the more important reasons is that:

--There are serious philosophic concerns about too much emphasis on specifiying program outcomes in graduate education, and expecting all students to move in the same direction or to value similar scholarly and professional aspects of a field. Some faculty argue persuasively that a major strength of advanced study is its flexibility and openness to intellectual idiosyncrasy. Partly because of such philosophic considerations:

--Student performance evaluation practices are characterized by an almost bewildering diversity in graduate education, with even such assessment staples as course grades and comprehensive examinations varying from one program to another in purpose, form, timing, and use. Such diversity obviously places both conceptual and practical constraints on any large scale research and development effort that would require standard practices, and suggests that attempts to explore new approaches might be much more likely to succeed if conceived and implemented as "local" rather than "central" activities.

--One important weakness of current evaluation practices is that often very little attention is apparently given to the purpose of evaluation. Many departments and programs evidently employ certain evaluation practices more for reasons of tradition than as a result of a careful consideration of what competencies are being assessed and why. Related to this problem, is the observation that:

--Performance assessment practices are seldom fashioned out of regard for different departmental training objectives or anticipated post-Ph.D. activities. Evaluations tend to focus on research competence, even though increasing numbers of those who earn the doctorate will spend much of their professional lives in teaching and other practitioner activities.

We emphasize here again, as we have elsewhere in the report, that these are observations which appear to reflect the general state of current student assessment practices in graduate education. There are certainly departments whose practices are clear and refreshing exceptions to the rule. In general, however, evaluation of student performance often seems to be taken for granted, and faculty members typically approach evaluation lacking any training in assessment and unclear as to why sound evaluation practices are important to the students, the department, or the profession. To the extent that this is an accurate analysis it does suggest some of the steps that might be taken at the departmental or institutional level to improve the evaluation of students.

What can Universities and Departments Do?

There are a number of steps that could be taken by institutions that would strengthen student performance assessment. Some of these would be more effectively carried out through the graduate dean′s office, others might be more appropriately located in the department.

(1) First, greater attention should be given to improving faculty members′ awareness of the importance of sound evaluation practices. Such a case is not difficult to make. Concerns about equity to students, maintenance of standards within the discipline, the quality of the curriculum and learning process--these and other considerations need to be emphasized. Useful activities could range from workshops to special training sessions for junior faculty to internal newsletters, and so on. The important point is that faculty members need to be convinced--and periodically reminded--that student performance evaluation is important for everyone concerned.

(2) Related to the first point is the observation that there is much that can be done simply by way of improving certain technical evaluation skills of faculty members. Some of these suggestions for improvement have been mentioned in various places throughout this paper.

(3) There should be periodic attention to the question of purpose, i.e., questioning of why certain forms of evaluation exist, what their strengths and weaknesses are, and whether different practices or forms might be better. Such analyses need to consider the goals of the program and objectives of the students.

(4) Procedures should be developed for keeping better records about various important aspects of student performance. Such records permit more careful tracking of student progress, and studying such information over time makes it possible to examine departmental standards and quality of student attainment.

(5) There needs to be an improved understanding of the problems of drawing inferences from validity studies based on small samples or weak criteria, and a greater emphasis on the rational basis for selection and performance measures used. For example, departments should undertake careful analysis of the skills and accomplishments they feel are especially relevant to their programs and use selection procedures which are geared to those factors. Local empirical studies are useful but not necessarily essential. Similarly, promising performance criteria should not be dismissed because of modest correlations with traditional predictors. The first emphasis should be on the employment of performance criteria that make good sense on the basis of the program's objectives and, accordingly, necessary student competencies.

(6) After being carefully spelled out, the basic aspects of the performance evaluation process should be routinely communicated to students. Students should be made aware of the nature and purpose of each aspect of the evaluation process, and entitled to an unambiguous report of their degree status and the department's appraisal of their performance at clearly specified points during graduate study.

Cooperative Research and Development Possibilities

Whereas the preceding suggestions all refer to steps that might be taken within single universities, there are numerous other activities that would seem to be appropriate and useful to consider at a different level within the graduate education community--policies or practices that could be encouraged and supported by graduate deans and various graduate education associations or groups, for example. Most of these efforts would require cooperation across graduate programs. Thus, the responsibility and primary initiatives might appropriately be in the hands of such organizations as the Graduate Record Examinations Board, the Council of Graduate Schools, the National Research Council, and the like. By giving encouragement, promotion, and both "official" as well as possible financial support, such agencies and organizations might:

(1) Hold national forums, special workshops, symposia at national disciplinary meetings and otherwise add their influence and assistance toward increasing awareness of the importance of sound student evaluation.

(2) Arrange for the collection and analysis of updated information regarding doctoral level attrition. As indicated earlier, no large-scale study of either the extent or causes of attrition has been conducted for more than a decade. Such a study could look not only at statistical information regarding attrition, but might well focus initially on case studies of departments that have developed effective ways of studying the attrition problem.

(3) Coordinate efforts that might be helpful in improving procedures for collecting criterion information in particular disciplines, especially with regard to student professional accomplishments. The routine and systematic collection of such information (and perhaps the sharing of that same information across a small number of collaborating institutions) would be a valuable educational experiment that might be quite helpful in clarifying the relation of training objectives to subsequent career roles and demands.

(4) Explore the feasibility of further developmental work on standardized work samples in the assessment of student performance along the lines described on pages 34-37. While such a project may not be practical as a large-scale activity, a small pilot examination of such work samples with a few cooperating departments could serve useful demonstration purposes to broaden the view of appropriate criteria and modes of assessment.

(5) Undertake efforts to work with interested institutions in an experimental project to examine the reliability and validity of the comprehensive examination. Such examina-

tions seem to vary widely in quality, but there is almost no literature concerning their character or principles of good practice in their development. Descriptions of a few exemplary comprehensive examinations could prove helpful to many departments.

(6) Consider the possibility of conducting (or sponsoring) a survey of student performance evaluation practices in America's graduate schools. Such a study might obtain useful summary information about the frequency of various evaluation practices, along with a clearer understanding of their purposes and uses.

The preceding observations and recommendations are based on our analysis of a variety of separate criteria used in the assessment of graduate student performance. The evaluation system is more than a sum of its parts, and the overall evaluation of graduate student performance is probably reasonably accurate and fair. In effect, a system of checks and balances operates in which, over a long period of time, shortcomings of one criterion are often offset by strengths in another. Thus, by using multiple--and often compensating--criteria, the failings in any one assessment are much less likely to lead to incorrect judgments about which students eventually earn the degree.

In addition, we need to be mindful of the sometimes-subtle nature of the graduate student evaluation process. We wrote earlier of how certain students who are judged to be inadequate are sometimes "cooled-out" of graduate programs, in spite of what would appear to be acceptable performance. Through various informal means, the faculty often maintains quality controls and discourages less able students, even though it may not be evident in the "public" records.

On the other hand, it is clear that certain specific criteria do have shortcomings and, could be substantially improved. It is our hope that this review contribute to a clearer understanding of what some of these problems are, and thus serves as a useful first step in making the procedures and criteria more helpful to both students and institutions.

References

American Psychological Association. Standards for educational and psychological tests. Washington, D. C.: Author, 1974.

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Baird, L. L. Using self-reports to predict student performance (Research Monograph Number 7). New York: The College Board, 1976.

Baird, L. L. Brief assessment: The validity and utility of biographical information and similar brief self-report measures. New York: College Entrance Examination Board, 1976a.

Baird, L. L. Development of an inventory of documented accomplishments: Report of Phase I and Proposal for Phase II, GRE No. 77-3, Educational Testing Service, December, 1976b.

Baird, L. L. Final report on Phase II of the project to develop an inventory of documented accomplishments. Princeton, N. J.: Educational Testing Service, August 1978 (unpublished draft for GREB).

Baird, L. L., & Richards, J. M., Jr. The effects of selecting college students by various kinds of high school achievements, ACT Research Report No. 23, Iowa City: American College Testing Program, 1968.

Barritt, L. S. The consistency of first semester college grade-point average. Journal of Educational Measurement, 1966, 3, 261-262.

Barzun, J. The American University. New York: Harper and Row, 1968.

Benson, W. W. Graduate grading systems. Proceedings of the Ninth Annual Meeting of the Council of Graduate Schools in the United States, Washington, D.C., 1969, ED 036 262.

Berelson, B. Graduate education in the United States. New York: McGraw-Hill, 1960.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. Taxonomy of Educational Objectives, Handbook I: Cognitive Domain. New York: McKay & Co., 1956.

Boldt, R. F. Factor analysis of business school grades. Research Bulletin RB-70-49. Princeton, N.J.: Educational Testing Service, 1970.

Bowers, J. E. A test of variation in grading standards. Educational and Psychological Measurement, 1967, 27, 429-430.

Breneman, D. W. Efficiency in graduate education: An attempted reform. Unpublished report to the Ford Foundation, 1977.

Bretsch, H. A study of doctoral recipients, 1938-1958. University of Michigan Graduate Study No. 6. Ann Arbor, 1965 (mimeo).

Brogden, H. E., & Taylor, E. K. The theory and classification of criterion bias. Educational and Psychological Measurement, 1950, 10, 158-186.

Bryant, N. D. A factor analysis of the report of officer effectiveness. Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center, 1956.

Caldwell, E., & Hartnett, R. T. Sex bias in college grading. Journal of Educational Measurement, 4, 3, Fall, 1967.

Carlson, A. B., Evans, F. R., & Kuykendall, N. J. The feasibility of common criterion validity studies of the GRE. Princeton, NJ: Educational Testing Service, 1973 (Research Memorandum 73-16).

Carlson, A. B., Reilly, R. R., Mahoney, M. H., & Casserly, P. L. The development and pilot testing of criterion rating scales, Princeton, NJ: Educational Testing Service, 1976.

Carmichael, O. C. Graduate education: A critique and a program. New York: Harper, 1961.

Carver, R. P. Two dimensions of tests: Psychometric and edumetric. American Psychologist, 1974, 29, 512 518.

Clark, D. C. Competition for grades and graduate student performance. Journal of Educational Research, 1969, 62, 351-354.

Clark, E. L. Reliability of grade-point averages. *The Journal of Educational Research*, 1964, *57*, 428-430.

Clark, M. J., Hartnett, R. T., & Baird, L. L. *Assessing dimensions of quality in doctoral education: A technical report of a national study in three fields*. Princeton, N. J.: Educational Testing Service, 1976.

Conrad, L. *Aptitude test restructuring research: Findings concerning projected development of an abstract reasoning measure*. Princeton, NJ: Educational Testing Service, September, 1976 (unpublished technical report for the Graduate Record Examinations Board).

Conrad, L., Trismen, D., & Miller, R. (Eds.). *Graduate Record Examinations Technical Manual*. Princeton, N. J.: Educational Testing Service, 1977.

Cronbach, L. J. *Essentials of psychological testing* (3rd edition), New York: Harper and Row, 1970.

Dale, W. Concerning grading and other forms of student evaluation. Proceedings of the Ninth Annual Meeting of the Council of Graduate Schools in the United States, Washington, D.C., 1969, ED 036 262.

Davis, J. A. What college teachers value in students. *College Board Review*, 1965, *56*, 15-18.

Doermann, H. Baccalaureate origins and performance of students in the Harvard Graduate School of Arts and Sciences. Unpublished report, 1968.

Donlon, T. F., Reilly, R. R., & McKee, J. D. *Development of a test of global vs. articulated thinking: The Figure Location Test* (GREB Report 74-9p). Princeton, N. J.: Educational Testing Service, June 1978.

Dressel, P. L., & Mayhew, L. B. *General education: Explorations in evaluation*. Washington, D.C.: American Council on Education, 1954.

Dyer, H. S. A review of General Education: Explorations in Evaluation, *Educational and Psychological Measurement*, 1956, 16, 153-160.

Eisenberg, T. Are doctoral comprehensive examinations necessary? *American Psychologist*, 1965, *20*, 168-169.

Frederiksen, N. There ought to be a law. Address presented at the ETS Invitational Conference on Testing Problems, October 29, 1977.

Frederiksen, N., & Ward, W. C. Measures for the study of creativity in scientific problem-solving. *Applied Psychological Measurement*, Winter 1978, vol. 2, no. 1, pp. 1-24.

French, J. W. *The description of aptitude and achievement tests in terms of rotated factors* (Psychometric Monograph No. 5), Chicago, Ill.: University of Chicago Press, 1951.

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.), *Psychological Principles in System Development*, New York: Holt, Rinehart, and Winston, 1962.

Goldman, R. D., & Slaughter, R. E. Why college grade-point average is difficult to predict. *Journal of Educational Psychology*, 68, 1, 9-14, 1976.

Gulliksen, H. *When high validity may indicate a faulty criterion* (ETS Research Memorandum 76-10). Princeton, N. J.: Educational Testing Service, 1976, 47 pp.

Hackman, J. R., Wiggins, N., & Bass, A. R. Prediction of long term success in doctoral work in psychology. *Educational and Psychological Measurement*, 1970, *20*, 365-374.

Halleck, S. L. Emotional problems of the graduate student. In Joseph Katz and Rodney T. Hartnett (Eds.), *Scholars in the making*, Boston: Ballinger, 1976.

Heine, R. W. Comparative performance of doctoral students admitted on the basis of traditional and non-traditional criteria. Unpublished manuscript, department of psychology, University of Michigan, 1976.

Heiss, A. *Challenges to graduate schools*. San Francisco: Jossey-Bass, 1970.

Hilton, T. L., Kendall, L. M., & Sprecher, T. B. Performance criteria in graduate business study. Parts I and II: Development of rating scales, background data and pilot study. (Research Bulletin 70-3), Princeton, NJ: Educational Testing Service, 1970.

Hirschberg, N., & Itkin, S. Graduate student success in psychology. American Psychologist, 33, 12, 1083-1093, December, 1978.

Holland, J. L., & Nichols, R. C. Prediction of academic and extracurricular achievement in college. Journal of Educational Psychology, 1964, 55, 55-65.

Humphreys, L. G. The fleeting nature of the prediction of college academic success. Journal of Educational Psychology, 1968, 59, 375-380.

Inbar, M., & Stoll, C. S. Simulation and gaming in social science. New York: The Free Press, 1972.

Juola, A. E. Freshman level ability tests versus cumulative grades in the prediction of successive terms performance in college. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, February, 1964.

Juola, A. E. Illustrative problems in college level grading. Personnel and Guidance Journal, 47, September, 196B.

Katz, J., & Hartnett, R. T. Scholars in the making. Cambridge, Mass: Ballinger, 1976.

Kelley, E. L., & Fiske, D. W. The prediction of performance in clinical psychology. Ann Arbor, Michigan: University of Michigan Press, 1951.

Lerner, B. The Supreme Court and the APA, AERA, NCME test standards: Past references and future possibilities. American Psychologist, October 1978, vol. 33, no. 10, pp. 915-919.

Lewy, A., & McGuire, C. A study of alternative approaches in estimating the reliability of unconventional tests, paper read at the annual meeting of the American Education Research Association, Chicago, February, 1966.

Manning, W. H. Test validation and EEOC requirements: Where we stand. Personnel, May-June, 1978, pp. 70-77.

Mayhew, L. R., & Ford, P. J. Reform in graduate and professional Education. San Francisco: Jossey-Bass, 1974.

McClelland, D. C. Testing for competence rather than for "intelligence." American Psychologist, 21, 1, January, 1973.

McGuire, C. H., & Babbott, D. Simulation technique in the measurement of problem-solving skills. Journal of Educational Measurement, 4, 1, Spring, 1967 (pp. 1-11).

Medsker, L. L., & Wattenbarger, J. L. An analysis of dissertations, 1975, Walden University, Mimeographed paper, 1976.

Meeth, L. R., & Wattenbarger, J. L. Dissertation quality at Walden University. Mimeographed paper, 1974.

Mooney, J. D. Attrition among Ph.D. candidates: An analysis of a cohort of recent Woodrow Wilson Fellows. Unpublished manuscript, Princeton University, 1967.

National Academy of Sciences. Doctorate recipients from United States universities, 1958-1966. Washington, D.C., 1967, 262 pp.

Nichols, R. C., & Holland, J. L. Prediction of the first-year college performance of high aptitude students. Psychological Monographs, 1963, 77, 7 (Whole No. 570).

Parry, S. B. The name of the game is simulation. Training and Development Journal, 1971, 25, 28-32.

Porter, Alan L., & Wolfle, Dael. Utility of the doctoral dissertation. American Psychologist, 30, 11, November, 1975, pp. 1054-1061.

Reilly, R. R. Critical incidents of graduate student performance. Princeton, NJ: Educational Testing Service, 1974a.

Reilly, R. R. Factors in graduate student performance. Princeton, NJ: Educational Testing Service, 1974b (Research Bulletin 74-2).

Richards, J. M., Jr., Holland, J. L., & Lutz, S. W. Prediction of student accomplishment in college. *Journal of Educational Psychology*, 1967, *58*, 5, pp. 343-355.

Rimoldi, H. J. Rationale and application of the test of diagnostic skills. *Journal of Medical Education*, 1963, *38*, 364-373.

Rosenhaupt, H. *Graduate Students′ experience at Columbia University, 1940-1956*. New York: Columbia University Press, 1958.

Singer, J. E. The use of manipulative strategies: Machiavellianism and attractiveness. *Sociometry*, 1964, *27*, 128-150.

Smith, G. M. Usefulness of peer ratings of personality in educational research. *Educational and Psychological Measurement*, 1967, *27*, 967-984.

Sparks, D. S. Grading and student evaluation. Proceedings of the Ninth Annual Meeting of the Council of Graduate Schools in the United States, Washington, D.C., 1969, ED 036 262.

Spurr, S. H. *Academic degree structures: Innovative approaches*. Report prepared for the Carnegie Commission on Higher Education. New York: McGraw-Hill, 1970.

Tansey, P. J., & Unwin, D. *Simulation and gaming in education*. London: Methuen Education Lts., 1969.

Thorndike, R. L. *Personnel selection: Test and measurement techniques*. New York: John Wiley & Sons, Inc., 1949.

Tucker, A., Gottlieb, D., & Pease, J. *Factors related to attrition among doctoral students*. Cooperative Research Project No. 1146, U.S.O.E., 1964

Tupes, E. C. *Personality traits related to effectiveness of junior and senior air force officers*, Lackland Air Force Base, Texas: Air Force Personnel Training and Research Center, 1957.

Tupes, E. C. *Relationships between behavior trait ratings by peers and later officer performance of USAF officer candidate school graduates*, Lackland Air Force Base, Texas: Air Force Personnel Training and Research Center, 1957.

Wallach, M. A. Psychology of talent and graduate education, in Samuel Messick and Associates, Individuality in Learning. San Francisco: Jossey-Bass, 1976, pp. 178-210.

Ward, W. C., & Frederiksen, N. A study of the predictive validity of the tests of scientific thinking. Princeton, NJ: Educational Testing Service, 1977 (Research Bulletin, 77-6).

Warren, J. W. College grading practices: An overview. ERIC Report No. 9, 1970.

Washington v. Davis. 426 U.S. 229 (1976).

Watson, G., & Glaser, E. M. Watson-Glaser critical thinking appraisal. New York: Harcourt, Brace, and World, 1964.

Wiggins, N., & Blackburn, M. Prediction of first-year graduate success in psychology: Peer ratings. Journal of Educational Research, 1969, 63, 81-85.

Wildberger, A. M. Review of modeling and simulation in education and training. AEDS Journal, 1974, 7, 65-74.

Willingham, W. W. Predicting success in graduate education. Science, 183, 273-278, January, 1974.

Willingham, W. W. Validity and the Graduate Record Examinations program. Princeton, N. J.: Educational Testing Service, 1977.

Willingham, W. W., & Breland, H. M. The status of selective admissions, in Selective Admissions in Higher Education (A Report of the Carnegie Council on Policy Studies in Higher Education). San Francisco: Jossey-Bass, 1977, pp. 66-244.

Wilson, K. M. Internal progress report of the Graduate Record Examinations Board Cooperative Validity Studies Project. Princeton, NJ: Educational Testing Service, 1978.

Wilson, K. M. Cooperative validity studies project: Testing a cooperative testing strategy (GREB 75-8). Princeton, N. J.: Educational Testing Service, in preparation.

Wilson, K. M. Of time and the doctorate. Atlanta, GA: Southern Regional Education Board, 1965.

Wing, C. W., & Wallach, M. A. College admissions and the psychology of talent. New York: Holt, Rinehart and Winston, 1971.

Wright, B. The Ph.D. stretch-out. Vital Issues in Education, 1957.