The Criterion Problem: What Measure of Success in Graduate Education?

Rodney T. Hartnett and Warren W. Willingham
Educational Testing Service

A wide variety of potential indicators of graduate student performance are reviewed. Based on a scrutiny of relevant research literature and experience with recent and current research projects, 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.

In any educational program a primary question is how to define successful performance. The so-called "criterion problem" has always been an important issue in validating admissions measures; what constitutes success also has a critical bearing on the very conception of a program and its objectives. 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 graduate student performance and how it ought to be measured naturally vary widely across institutions, disciplines, and types of programs. As a result, there is often ambiguity in the meaning of "success" in graduate school, and a corresponding set of issues and questions that must be addressed when embarking on research—especially validity studies research—that relies heavily on graduate school performance as a criterion. Therefore, an overview of the criterion problem as it applies to graduate education would seem to be much overdue. This review distinguishes three broad classes of criterion measures: traditional criteria (e.g., grades, examination performance), evidence of professional accomplishment (e.g., publications, awards), and specially developed criteria (e.g., work samples, ratings).

Traditional Criteria

A number of criteria have been used for years in the assessment of graduate student performance. These 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.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 3 Summer 1980 pp. 281–291

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 (GRE; Willingham, 1974).

As an indicator of student performance, grades have several positive qualities. First, 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. In addition, grade-point averages 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 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, and level of curiosity—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, 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 under-graduate 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 (Humphreys, 1968; Juola, 1964). At the graduate level, evidence regarding the stability of grades is more difficult to find. It is clear that there is less fluctuation 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. 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. Grades at the graduate level thus may not provide meaningful descriptions of differential student performance.

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 & 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 (nonpooled) analyses, is positive.

A third difficulty with grades as a criterion is that it is not always clear what grades mean. Dif-

ferent 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), gender (Caldwell & Hartnett, 1967), or various manipulative strategies (Sanford, 1976). 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, & Bass, 1970), and it is likely that the basis for grading is quite different before and after students are accepted to formal candidacy.

Degree Attainment

Degree attainment has been employed in validity studies as often as the grade-point average (Willingham, 1974) and is a useful and important criterion of graduate student performance. 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—such as grades or faculty ratings—are simply poor proxies for what really counts: namely, whether the student eventually earned his or her degree. Graduate students clearly regard it as the most important outcome of their graduate studies.

As a criterion for the study of graduate student performance, however, degree attainment has certain limitations. One limitation is that 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, however, may never be learned. As Berelson (1960) has pointed out, there is sometimes—how frequently we cannot say—a discrepancy between the real reason and the reason reported by those withdrawing. According to Berelson, "What is critical frankness between doctoral candidates becomes in the dean's office lack of funds or personal change of plans" (p. 170).

Another shortcoming of degree attainment as a criterion is the fact that most graduate programs keep very inadequate records about attrition (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 *indecision* 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.

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 has received considerable attention from researchers, both at the national level (National Academy of Sciences, 1967; Tucker, Gottlieb, & Pease, 1964; Wilson, 1965); and at various doctoral-granting universities such as Columbia (Rosenhaupt, 1958), Michigan (Bretsch, 1965; Heine, 1976), and Harvard (Doermann, 1968).

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 do not earn the degree sooner because they can not, 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. 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 other nonacademic factors (Katz & Hartnett, 1976). Berelson (1960) even suggests that some students actually are not allowed to finish sooner because ". . . they are needed as teaching assistants for the department or as research assistants for the professor" (p. 162).

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, the terms "comprehensive examinations" and "qualifying examination" will be used interchangeably.)

One of the most commonly criticized weaknesses of comprehensive examinations 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. There is some evidence to suggest that some departments do not take the comprehensive examinations very seriously and are apparently not very concerned about taking steps that would make them more reliable and meaningful measures of student attainments (Berelson, 1960; Heiss, 1970; Mayhew & Ford, 1974). And, 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—there is no evidence for this assertion—that evaluations of student comprehensive examination performance are often not very reliable.

In spite of these shortcomings, many graduate faculty members appear to be reluctant to consider more systematic procedures for the assessment of student academic attainment (Carlson, Evans, & Kuykendall, 1973). 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, in spite of the success

of such practices in several professional fields (e.g., McGuire & Babbott, 1967; Rimoldi, 1963).

Dissertation Quality

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 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 nonresident doctoral programs (e.g., Medsker & Wattenbarger, 1976; Meeth & Wattenbarger, 1974), there has been scant attention given to the dissertation as an indicator of 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.

On the other hand, it is clear that numerous practical difficulties would be encountered in any attempt to employ dissertation quality as a criterion of research competence. For example, 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 to make ratings on standard, carefully constructed dimensions of quality. Thus, any such undertaking would be fairly expensive and time consuming. 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 hypothesis of the inquiry, the basic research design or strategy, or the 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% 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.

Informal Criteria

One final observation is in order before the review of administrative criteria is concluded. 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 cannot be said—operates in a more informal fashion and its results never become part of any formal record. For example, many faculty members 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 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 (Katz & Hartnett, 1976; Sanford, 1976). 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 that 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).

Most of these efforts have concentrated on the prediction of undergraduate performance on the basis of secondary school (or nonschool) accomplishments. Recently, however, Baird (1976) developed an experimental inventory of undergraduate accomplishments that might be used in graduate school admissions. 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.

Such indices of professional behavior have considerable merit as criteria because they reflect important long-term objectives. However, 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, there would be several significant limitations with student professional accomplishments. One is that such accomplishments may be partly a matter of simple luck. 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 this regard, 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 or a journal article with one professor, an even more substantial contribution from another student may not even earn an "I am indebted" footnote. 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 that the distribution of such accomplishments will be extremely narrow and skewed. At least this is true at the undergraduate level (Baird, 1978), and it is almost surely the case in graduate school as well. This

does not affect the logic of using accomplishments as a criterion, of course, but it does reduce their likely utility.

Specially Constructed Criteria

In addition to traditional criteria and student professional 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 two types of constructed criteria are considered: rating scales and performance work samples.

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 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 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) 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 that 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/her 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, 1957a, 1957b); more recently, their potential in educational settings was suggested when it was found that peer ratings of nonintellective 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.

Performance Work Samples

Frederiksen (1977) has argued that before developing a measure of the effectiveness of a training program, the kinds of skills and behaviors to be expected from those who have experienced the training need to be understood first; 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 a clearer understanding of the specific behaviors and activities that will be expected of the graduates of most graduate programs is needed. These people, after all, will be employed in positions requiring very complex behaviors and skills, ones neither 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 minute detailed 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).

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 to 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, as suggested above. 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 has resulted in the development of The Tests of Scientific Thinking (TST), which are free-response job-sample tests simulating tasks that might be encountered by a

behavioral scientist (Frederiksen & Ward, 1978; Ward & Frederiksen, 1977). Research with the experimental TST has indicated that the various TST subtests 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.

As with the other performance criteria, performance work samples have a number of limitations. For one thing, 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 in a branch of a discipline. Because of the apprenticeship nature of the graduate experience, in certain disciplines there is often little substance in common among the various subspecialties. Another problem is that 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 is 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.

Conclusions

This paper has 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. Though the intention was to examine these criteria primarily in terms of their adequacy as dependent variables for the validation of graduate school selection procedures, it is apparent that each separate criterion can be fully understood only after consideration of the more general process of graduate student evaluation. As a result, though this review deals primarily with the criterion problem in the traditional measurement sense, it also, to some extent, provides an overview of student performance evaluation in graduate education.

Perhaps the most important observation is that, overall, 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. To some extent this lack of empirical attention can be explained by a concern, among graduate faculty members, about too much emphasis on specifying program outcomes. Some argue persuasively that the major strength of advanced study is its flexibility and openness to intellectual idiosyncrasy. This view no doubt explains, at least in part, why 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.

These general observations, along with the more detailed shortcomings of specific criteria discussed earlier in the paper, suggest that the criterion problem can be expected to continue to be bothersome to those conducting research on graduate student performance. Attention to certain psychometric characteristics of standard criteria (such as improving the inter-reader agreement on comprehensive examinations) or the willingness to consider the possible merits of new criteria (e.g., performance work samples, global ratings) may yield modest gains. But the nature of the measurement problems are so pronounced, and the logistic and philosophic realities so chronic, that for the foreseeable future measurement specialists will have to be content with less-than-satisfactory criterion measures when embarking on research on graduate student performance.

References

Baird, L. L. *Development of an inventory of documented accomplishments: Report of phase I and proposal for phase II* (GRE No. 77-3). Princeton, NJ: Educational Testing Service, December, 1976.

Baird, L. L. *Final report on phase II of the project to develop an inventory of documented accomplishments* (Unpublished draft for the Graduate Record Examination Board). Princeton, NJ: Educational Testing Service, August 1978.

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, IA: 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.

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

Boldt, R. F. *Factor analysis of business school grades* (Research Bulletin RB-70-49). Princeton, NJ: Educational Testing Service, 1970.

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

Bretsch, H. *A study of doctoral recipients, 1938-1958* (Graduate Study No. 6). Ann Arbor: University of Michigan, 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, TX: Air Force Personnel and Training Research Center, 1956.

Caldwell, E., & Hartnett, R. T. Sex bias in college grading. *Journal of Educational Measurement*, 1967, *4*, 129-133.

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

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.

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, NJ: Educational Testing Service, 1976.

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

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.

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 1977.

Frederiksen, N., & Ward, W. C. Measures for the study of creativity in scientific problem-solving. *Applied Psychological Measurement*, 1978, *2*, 1-24.

French, J. W. The description of aptitude and achievement tests in terms of rotated factors. *Psychometric Monographs*, 1951 (No. 5).

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagné (Ed.), *Psychological principles in system development*. New York: Holt, Rinehart, & Winston, 1962.

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

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 J. Katz & R. T. Hartnett (Eds.), *Scholars in the making*. Cambridge, MA: Ballinger, 1976.

Heine, R. W. *Comparative performance of doctoral students admitted on the basis of traditional and non-traditional criteria*. Unpublished manuscript, 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*, 1978, *33*, 1083-1093.

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.

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,* 1968, *47,* 29-33.

Katz, J., & Hartnett, R. *Scholars in the making.* Cambridge, MA: Ballinger, 1976.

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

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,* 1973, *28,* 1-14.

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

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

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

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

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).

Porter, A. L., & Wolfle, D. Utility of the doctoral dissertation. *American Psychologist,* 1975, *30,* 1054-1061.

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

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

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

Sanford, M. *Making it in graduate school.* Berkeley:Montaigne, 1976.

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.

Tucker, A., Gottlieb, D., & Pease, J. *Factors related to attrition among doctoral students* (Cooperative Research Project No. 1146). Washington, DC: U.S. Office of Education, 1964.

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

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

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

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

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

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

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

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

Acknowledgment

This research was supported by a grant from the Graduate Record Examinations Board.

Author's Address

Send requests for reprints or further information to Rodney T. Hartnett, Senior Research Psychologist, Higher Education Research Group, Educational Testing Service, Princeton, NJ 08541.