DOCUMENT RESUME

ED 068 577 TM 002 092

AUTHOR	Quirk, Thomas J.
TITLE	A Manual for Designing and Conducting Validity Studies Based on the National Teacher Examinations.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	PR-72-8
PUB DATE	May 72
NOTE	54p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Certification; Computer Programs; Correlation; Data Collection; *Evaluation Criteria; Guides; Higher Education; *National Competency Tests; Standardized Tests; *Teacher Certification; *Test Construction; *Validity
IDENTIFIERS	*National Teacher Examinations; Validity Study Service

ABSTRACT

A detailed outline is provided of the steps necessary to use the Validity Study Service provided by the National Teacher Examinations (NTE) program at Educational Testing Service. It is intended to assist those school districts, teacher-training institutions, and state certification offices that have NTE data to establish a standard system for data collection, to allow for the analysis of the data by means of a common set of computer programs, and to assist the test users in the development of local norms and local correlational studies. The use of testing jargon is minimal, and the concepts that are presented in technical terms are supplemented by listings in the glossary of key terms. Section I of the manual presents a brief discussion of test validity. Section II discusses ways of selecting the sample of candidates to be studied, the choice of predictors and criteria, and the necessity for cross-validation. Section III describes the report that is sent to the institutions who use the computer programs on which this manual is based. Section IV then describes the necessary data collection and coding procedures. (LH)

PR-72-8

A MANUAL FOR DESIGNING AND CONDUCTING VALIDITY STUDIES BASED ON THE NATIONAL TEACHER EXAMINATIONS

Thomas J. Quirk

May 1972

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY

A MANUAL FOR DESIGNING AND CONDUCTING VALIDITY STUDIES BASED ON THE NATIONAL TEACHER EXAMINATIONS

Thomas J. Quirk
Teacher Behavior Research Group
Educational Testing Service

CONTENTS

PREFACE

This manual provides a detailed outline of the steps necessary to use the Validity Study Service provided by the National Teacher Examinations (NTE) program at Educational Testing Service. It is intended to assist those school districts, teacher-training institutions, and state certification offices that have NTE data to establish a standard system for data collection, to allow for the analysis of the data by means of a common set of computer programs, and to assist the test users in the development of local norms and local correlational studies.

In writing this, I have not attempted to offer a compressed course in tests and measurements or even the concepts of test validity. A definitive article on test validity has already been written by Cronbach (3), and a simpler version of some of the basic concepts of that article can be found in Cronbach and Quirk (4). The reader who is interested in the problem of test validation will learn more by reading those two sources than he will from this manual. I have tried, whenever possible, to minimize the use of testing jargon. There are some concepts, however, that are more efficiently presented by means of technical terms, and whenever such terms are used, they are defined in the alphabetical Glossary of Key Terms.

Section I of the manual presents a brief discussion of test validity. Section II discusses ways of selecting the sample of candidates to be studied, the choice of predictors and criteria, and the necessity for cross-validition. Section III describes the report that is sent to the institutions who use the computer programs on which this manual is based. Section IV describes the necessary data collection and coding procedures.

Several colleagues at Educational Testing Service have provided valuable assistance in tracing the myriad details necessary to write this manual; in particular, I should like to thank Betty Humphry, Eleanor Weiss, Arthur Benson, Hadley Nesbitt, Richard Majetic, and Don Oppenheim for their patience and cooperation in helping me find the answers to a host of detailed questions. Nat Hartshorne was especially helpful in asking the kinds of questions about my writing style that improved this report considerably. I thank all of them.

T. J. Q.

I: TEST VALIDITY

The National Teacher Examinations and How They Help

The National Teacher Examinations (NTE) offered by Educational Testing Service (ETS) provide an independent assessment of the academic preparation of teacher-education candidates who are college seniors completing a four-year program in teacher education. The NTE are national, standardized, secure tests that permit comparison of candidates within the same institution and across different institutions within the limits defined by the test content. School districts use NTE scores as one part of their selection process for beginning teachers. Teacher-training institutions use the scores as one part of their evaluation process of teacher-training candidates. Educational Testing Service does not set any passing or failing standards for any of the National Teacher Examinations. Only local institutions can make this type of decision based on their own validity studies of the NTE.

The National Teacher Examinations are not intended to measure teacher aptitude, interests, attitudes, motivation, maturity, or other personal or social characteristics of beginning teachers. Nor are they intended to be a measure of classroom teaching performance. Educational Testing Service does not claim that NTE scores will predict teaching effectiveness. What a teacher knows about his teaching area of specialization may or may not indicate what he will do in the classroom.

The NTE currently provide achievement scores in the Common Examinations and in 24 Teaching Area Examinations. For a discussion of how the tests are planned and constructed, see the National Teacher Examinations: Interpretation of Scores (7), the Bulletin of Information for Candidates 1970-71 (8), the NTE Prospectus for School and College Officials (9).

Improper Uses of the NTE

NTE scores are not always used properly. For example, using them to rank teacher-training institutions in a state is not practical unless such a comparison includes the different standards of the institutions in admitting, training, and graduating teacher-education candidates. Sometimes a single NTE score is set as a cutoff, and all candidates who score below this level are automatically rejected for a teaching position. We strongly oppose the use of an inflexible cutoff score on the NTE. The tests are a less-than-perfect measure of a candidate's knowledge in general education and subject-matter competency. Moreover, test scores should be considered along with a wide variety of other information such as attitudes, personality, maturity, academic record, personal and social skills, recommendations for qualified personnel, teaching experience, out-of-school activities, and potential as a teacher.

To use the NTE as an entrance requirement for admission to graduate school, a purpose for which the test is not intended and for which norms are not available, may be inappropriate unless one can demonstrate a significant relationship between the NTE and graduate school performance.

Finally, the NTE scores should never be used in decisions about retention or tenure, and only with caution in hiring, experienced teachers. The NTE Guidelines for Using the National Teacher Examinations (10) have stated it well: "When an adequate and reliable record of the teacher's performance is available, there is no need to attempt to predict his teaching abilities." For a more detailed discussion of the uses and misuses of the NTE scores, see the Guidelines.

Test Validation

How do school administrators and other test users know when or if they are using NTE scores properly? The procedure by which this is studied is called test validity. The validation of a test requires an investigation of the accuracy of a prediction from a test score. Since there are many different ways of using a test with different candidates, there is no such thing as the validity of a test. How well a test carries out the purpose for which it is used determines whether it has any validity for that purpose.

Similarly, one cannot ask how high a validity correlation should be. To the test user, a test that predicts an outcome that is important to him with a correlation of .30 can be much more useful than another test that predicts what he considers to be a less important outcome with a correlation of .70. Tests that are moderately correlated with an outcome can be valuable if the number of candidates selected is very small compared with the number of applicants and when there are large differences between the candidates in the performance of the outcome.

Establishing a Selection Procedure

After administering the test, the first step in any validation process is one of selection by which some candidates are chosen to participate in a particular experience while others are rejected. The validation of an institution's selection procedures is most effective when the selection is not influenced by candidates' test scores. In this way, the test scores can be compared with existing sources of information to see if they distinguish successful from unsuccessful candidates in terms of subsequent performance.

The percentage of candidates who normally succeed when the test is not used for selection is called the base rate. Examples of the base rate in

teacher education would include the percentage of candidates who successfully complete a teacher-training program and the percentage of teacher-training students who complete at least one year's full-time teaching experience within three years of completion of the teacher-training program. If the base rate is very high so that virtually all applicants are successful in the teacher-training program or job, money spent on testing for the purpose of admission to the program or job may be wasted. Trying to weed out the 5 percent failures by a screening test, for example, may result in so many errors in prediction that it would be both simpler and less costly to accept everyone.

Establishing a selection procedure is not always simply a matter of finding candidates similar to employees who have been highly successful on the job. For example, in using a test to select teachers, it is not sufficient to show that experienced teachers who are judged to be outstanding are the ones who also score high on an instrument such as a general culture test. Perhaps prospective teachers who score low on such a test would be acceptable as teachers if they could only sneak past the employment office. Or perhaps these prospective teachers could learn quickly what they need to know by being trained on the job. A test can justifiably be used for selection only by showing that candidates who score low on the test turn out to be poor teachers even after a reasonable training effort has been made.

The Function of Predictors and Criteria

A predictor is a test or some other device used to estimate an individual's performance on some outcome, which is called the criterion. When one attempts to relate the rank order of candidates on the predictor to the rank order of these same candidates on the criterion, we speak of the predictive validity of the test. Table 1 lists predictors and criteria that have been or could be used in

research studies of teacher selection. Not all of the predictors listed in Table 1 can be defended; moreover, many can be used as criteria as well as predictors. The challenge from a research point of view is to select those that will be most useful.

Choosing predictors and criteria should be done carefully. The criteria, by definition, become the yardsticks against which the effectiveness of the candidate is judged, and, for this reason, should be comprehensive enough to include a variety of the complex behaviors required in full-time teaching. The criteria selected for use should be given the same scrutiny by the test user in terms of their reliability and validity as the predictors.

In selecting predictors, a general rule is to try to find ones that are work samples of what real, full-time teaching is like. The high school grade-point average is a good predictor of college grade-point average, since both averages often measure directly the same type of academic skills. But a work sample of full-time teaching is difficult to find. Experience in student teaching probably comes closest to being a work sample in most teacher-training programs, but grades given for that experience tend to be based on such a wide variety of factors, combined with such a different weighting system, that their meaning across student teachers, and certainly across different teacher-training institutions, is at best confusing; the problem is compounded by the fact that the grade in a single course is not likely to be highly reliable. If, in practice, almost all students enrolled in a teacher-training program receive a grade of A or B in student teaching, the spread of grades is so small that prediction of individual differences could be quite unreliable. Further, because a certain test predicts a grade in student teaching fairly well does not mean that such a test will predict success in full-time teaching equally well. Full-time teaching may be much more demanding in its complexity than

TABLE 1

A List of Possible Predictors and Possible Criteria for Teacher Behavior

Possible PREDICTORS	Possible CRITERIA
High school grade-point average (GPA)	
High school rank in class (converted score)	
College Entrance Examination Board Scholastic Aptitude Test	
College Entrance Examination Board Achievement Tests	
College-Level Examination Program	
Self-reports	Self-reports
Biographical Inventory	Biographical Inventory
Questionnaire	Questionnaire
Interest Inventory	Interest Inventory
Statisfaction with teacher training program	Satisfaction with teacher training program
Satisfaction with practice teaching	Satisfaction with practice teaching
Career plans and expectations	Career plans and expectations
Satisfaction with first year of full-time teaching experience	Satisfaction with first year of full-time teaching experience
Number of years of teaching experience	Number of years of teaching experience
College overall GPA	College overall GPA
GPA in academic major	GPA in academic major
GPA in education courses	GPA in education courses
Grade in practice teaching	Grade in practice teaching
Persistence (completion of training program)	Persistence (completion of training program)
National Teacher Examinations	National Teacher Examinations
Common Examinations	Common Examinations
Teaching Area Examinations	Teaching Area Examinations

Table 1, continued

Possible PREDICTORS	Possible CRITERIA
Ratings Pupils College field supervisor School district supervisor School principal Peers (other teachers) Faculty or counselor recommendation	Ratings Pupils College field supervisor School district supervisor School principal Peers (other teachers) Faculty or counselor recommendation
End-of-course examination	End-of-course examination
Departmental examination	Departmental examination
Interview information	Interview information
Personality measures	Personality measures
Graduate Record Examination scores	Graduate Record Examination scores
End-of-training scores	End-of-training scores
In-basket exercises	In-basket exercises
Simulation tests (e.g. micro-teaching tests)	Simulation tests (micro-teaching tests)
	Job Performance scores
	Classroom observation by trained observers (time-sampling)
	Completion of first year of full-time teaching within three years of completion of the teacher-training program (persistence)
	Number years of teaching in the district
	Total number of years of teaching experience (full-time equivalent)
	Number of promotions (department head, assistant principal, principal, etc.)
	Teacher transfer out of school district
	Pupils' achievement test scores
	Performance tests

the typical student teaching experience. This is one of the reasons why long-term follow-up studies of graduates of teacher-training programs are an essential aspect of research studies designed to check on the effectiveness of such programs.

Ratings of Teachers by School Administrators

The usefulness and accuracy of ratings of teacher performance have been questioned for many years. As Cronbach has put it (2): "When a test fails to predict a rating, it is hard to say whether this is the fault of the test or of the rating." Ratings can easily reflect the degree to which the rater likes the teacher rather than the quality of the teacher's work. In some cases, the rater may simply not know the facts about the teacher. Stories of teachers who claim that they were rated by someone who visited their classes a total of only 15 to 20 minutes during the entire school year are common in teachers' lunchrooms. This small sampling of the classroom behavior of the teachers can hardly be considered adequate.

Raters attach different meanings to the traits on which they rate teachers. "Leadership" might mean relying on authority, dominance, and clear decision-making to one rater and encouraging pupils, working out cooperative decisions between teachers and pupils, and democratic rule to another. Moreover, the rating scale itself may be ambiguous. Rating "cooperativeness," "adaptability," or "sensitivity" on a scale from 0 to 10, or from poor to excellent, is hopeless unless clear descriptions are given for each point on the scale in terms of actual teacher behavior.

Ratings are most useful when raters can agree on what they see and on how they will code the teacher's behavior and when the teacher's behavior is

not likely to vary a great deal over time. The best way to obtain useful information from raters is to instruct them carefully about the definitions of the items, show them examples of actual teacher behavior for each item, and check the reliability of their ratings of actual classroom situations. It might also prove useful to use raters who do not know the teachers personally. School principals must convince school district superintendents that they are developing their teachers into outstanding members of the profession. Thus, they have a large personal stake in the ratings.

Correlation Coefficients

The summary of how well a predictor estimates a criterion is typically expressed in the form of some type of correlation coefficient. Most types of correlation coefficients range from -1.0 to +1.0; the higher the coefficient, the better the test can predict the outcome.

The range of individual differences of the candidates being studied, expressed typically as the standard deviation of scores, can reduce the magnitude of the correlation coefficient. If the standard deviation is smaller, the correlation will be lower. For example, if the standard deviation of the candidates in a research study is 100 and the correlation between predictor and criterion is .64, and in later years the standard deviation of the candidates reduces to 70, the correlation between predictor and criterion will be reduced to .50 even if nothing else changes (5).

The statistical model on which are based the computer programs that apply to the Validity Study Service outlined in this manual is the linear multiple-regression model. This model is based on the assumption that a higher degree of ability in one area of knowledge or skill can compensate for a lower ability in another by "averaging" those abilities considered

to be important by the test user. The procedure by which more than one predictor is used to estimate a criterion score is called multiple regression. Thus, two or more predictor scores would be weighted appropriately and combined in order to predict a criterion. An excellent discussion of multiple regression and the appropriate mathematical formulae can be found in Ghiselli's Theory of Psychological Measurement (5).

The correlation coefficient resulting from using two or more predictors to estimate a criterion is called a multiple-correlation coefficient. In order to obtain a high multiple-correlation coefficient, the predictor tests should have a high correlation with the criterion and low correlations with each other. The predictor test scores are combined to estimate the criterion score by weighting the predictors. These weights are determined mathematically and are less likely to vary drastically from one sample of candidates to another if the predictors measure different things, if the weights are derived from a large sample of candidates, and if the method of measuring the criterion does not change either in the method or the accuracy of assessment.

The multiple-correlation procedure tells us how much each predictor score adds to the prediction of the criterion by each predictor separately. A test should be included in a prediction equation only if it adds significant information to what other, readily available, simpler, and less expensive information can provide. This is another way of saying that the team of predictors in a multiple-regression equation must be considered as a set of predictors so that no single predictor can be considered independently of each of the others. A given predictor should be included in the multiple-correlation equation only if it increases the multiple-correlation coefficient sufficiently to justify the time and expense of collecting the information on the predictor in question.

The weights attached to each of the predictors in a multiple-correlation equation are derived from the particular sample of candidates studied. Thus, the magnitude of these weights would be expected to change from one sample of candidates to another by chance alone. In order to check on the magnitudes of the weights, the test user should apply the weights derived from one sample directly to another sample of similar candidates to see if the magnitude of the multiple-correlation coefficient "shrinks" appreciably in predicting the criterion score. A mathematical formula for estimating the shrinkage can be found in McNemar (6). A good procedure for checking the accuracy of the weights is to divide the sample of candidates randomly into two groups, derive the weights on one of the groups and apply them to the other group to see if the correlation holds up. This process -- called cross-validation -- is not a luxury but an essential part of any validity study involving multiple-correlation techniques.

Developing Local Validity Studies

School systems that use local norms and local validity studies can capitalize on their special, first-hand knowledge of the local situation and of the candidates from whom they must choose their teachers. A research study that predicts success in teaching is most applicable in the locale in which it was developed and when the local situation in which it was developed remains sufficiently stable that the findings remain relatively constant over time. Such components as the curriculum, faculty, and the quality of students are continually changing, and in order to predict the future, one must judge the similarity between past and present conditions. Cronbach (2) summarizes this well:

> However well a test has been developed and however thoroughly its author has validated it, no one can be sure it will predict in a situation until it is tried out there.... Sooner or later, nearly every person using tests for selection or classification must carry out his own validation studies to determine whether his prediction methods are working.

Sometimes data can be analyzed separately for different subgroups of candidates. Some of the more useful of these for teacher selection would be sex, age, amount of education, teaching-area speciality, level of training, and years of teaching experience.

A useful source of information for those planning to conduct local validity studies is entitled Constructing and Using Local Norms (1) and is available from Cooperative Tests and Services, Educational Testing Service, Princeton, New Jersey 08540.

II: THE DESIGN OF VALIDITY STUDIES

The design of any validity study should be planned to answer only those specific questions the test user is most interested in, and the data must be prepared in such a way that these questions can be answered adequately.

Selecting the Group to be Studied

The minimum size of a subgroup should be approximately 85 candidates whenever possible. This sample size should be thought of as a goal rather than a fixed standard; there will undoubtedly be many instances in which only smaller samples are available. If fewer than this number are available, the test user can pool candidates over the past year or two to reach the necessary minimum size. It does not make sense to pool candidates from different years if there have been drastic changes in selection policies, grading practices, curricula, or in the method of evaluating criterion performance during this period, since any of these changes would make the groups dissimilar. If possible, at least 85 percent of the candidates in each subgroup should have scores on all predictors and criteria in a validity study; a lower percentage is acceptable if there is reason to believe that there is no bias in the incomplete data group, but this judgment is usually a difficult one to make.

As the number of predictors increases, the size of the subgroups should also increase. In addition, any procedure for selecting candidates that would make the group less similar to future groups of interest to the test user should be avoided. For example, do not drop from the data analysis those candidates who have unusual scores or those who fail to complete the teacher-training program. Any such deletion would mean losing a very useful piece of information.

Choosing the Predictors

Each teacher-training institution, school district, or state certification office should designate, or at least form hypotheses about, those predictor measures that will be useful in the local situation and how they are related to each other and to the criteria. Only the test user knows his local situation well enough to select those predictors that interest him.

Data from interviews or ratings can be included in validity studies as long as this information is coded onto a numerical scale. Whenever possible, at least 90 percent of the candidates should have scores on each predictor. But candidates should not ordinarily be deleted from the group in order to reach this percentage; any such tampering might bias the original sample of candidates to an unknown degree.

Choosing the Criteria

Once the test user has chosen the subgroups and the predictors, the next step is to select the criteria that he is most interested in predicting. Criterion information should be comparable for every candidate in the groups studied; for example, it is improper to use a three-year grade-point average (GPA) for some candidates and a four-year GPA for others.

The criteria should be selected with caution. Some criterion measures do not make any sense; for example, using Graduate Record Examinations (GRE) scores as a criterion for a teaching credential would be inappropriate since the GRE Board makes no claim that their tests are valid or that they have norms for such a purpose.

Up to three criteria can be selected by the test user for each subgroup. Criterion scores must be expressed in quantitative form, and the underlying scale of measurement should be linear. Two-category classes, such as pass-

fail or successful-unsuccessful, can be included by coding one of these categories zero and the other as 1.

Using the NTE Common Examinations

It is quite proper to use the NTE Weighted Common Examinations Total score (WCET) in validity studies, since these scores have been equated from form to form since 1940.

Since the WCET score is the only score of the Common Examinations that has been equated from form to form, however, it is hazardous to use any of the other scores of the Common Examinations in validity studies. Thus, scores on Professional Education, General Education, Psychological Foundations of Education, Societal Foundations of Education, Teaching Principles and Practices, Written English Expression, Social Studies, Literature, and the Fine Arts, and Science and Mathematics should be used with caution as predictors or criteria in validity studies since the applicability to future studies of findings based on these scores will be restricted to an unknown degree.

Using the NTE Teaching Area Examinations

During 1971, 24 Teaching Area Examinations (TAE) were offered in the NTE testing program. The scaled scores for the TAE are based on substantially all candidates who indicated at the national administration in February 1964 that the TAE they took were in the field for which they were best prepared to teach. Since February 1964, each new form of each TAE has been equated statistically to earlier forms of the same TAE to allow for differences in the difficulty and length of subsequent test forms.

Scores on a TAE should be included in a validity study only if all candidates being compared took the TAE in the same subject-area speciality after February 1964. TAE scores prior to 1964 should never be used in validity studies.

Using an inflexible cutoff score across TAE is not advisable. It is not possible to say, for example, that a senior who scored 700 on the Mathematics TAE is a better candidate than a senior who scored 600 on the Early Childhood Education TAE, since these two seniors took different tests. Moreover, a cutoff score of 600 would eliminate 30 percent of the mathematics seniors, only 23 percent of the Early Childhood Education seniors, and only 15 percent of the Biology and General Science seniors.

Using the Composite NTE Score

A composite NTE Score is the sum of the WCET score and the TAE score. It is incorrect to use these three scores — the WCET score, the TAE score and the Composite NTE score -- as predictors in the same validity equation since, in this case, the composite NTE score would be the sum of two scores that are already included in the study. It is also incorrect to use the composite NTE score as a predictor unless all of the candidates in the group or subgroup being studied took the TAE in the same subject-area speciality and since February 1964.

It is possible to include a weighted composite score, such as WCET + 2 TAE as a predictor in a validity study as long as all of the candidates being compared took the TAE in the same subject-area speciality and after February 1964.

Using Cross-validation Procedures

New groups of candidates may differ from previous groups in systematic ways so that what was true in the past no longer applies. For this reason, validity studies should be repeated often so that assumptions can be reevaluated and decision rules updated.

Through the process of cross-validation, an institution takes the weights applied to predictor scores for one group of candidates and applies them to a second group of similar candidates in order to cross-check the effectiveness of the predictors in estimating a criterion. These weights, when applied to

the second group of candidates, produce an estimated criterion score for each candidate which can be compared with his actual criterion score. The correlation between the predicted criterion scores and the actual criterion scores can indicate the accuracy of the original predictor equation. Whenever possible, the weights derived from one sample of candidates should be cross-validated on another sample of similar candidates in order to check the accuracy of the prediction equation derived from the first.

III: THE DATA ANALYSIS REPORT

The Data Analysis Report, which is sent to those who conduct validity studies through ETS, contains the following information:

- A frequency distribution of predictor scores and criterion scores for each variable studied will show the number of candidates (N) who have scores within the intervals indicated by the limits of the observed scores and the percentage of candidates whose scores fall below the lowest score in each interval (percent below).
- An average score (mean score), a measure of the spread of scores about that average score (standard deviation), and the highest and lowest obtained score will be reported for each variable. The frequency distributions and the means and standard deviations can be used by the test user to develop local norms.
- The correlation between each predictor and the criterion will be reported in the form of a correlation table. The test user can examine these correlations to discover not only the best single predictor of the criterion but also to discover which predictors do not estimate the criterion very well. The standard error of estimate for each predictor and for each set of predictors will also be provided.
- The Summary of Statistics section of the report will indicate the correlation of each predictor with the criterion, the mean, standard deviation, and standard error of estimate for each predictor, the mean and standard deviation of the criterion, and the improvement of the prediction when two or more predictors are used to

estimate the criterion. The degree of improvement will be indicated by a correlation coefficient using the sets of predictors specified by the test user. If the test user decided from the data analysis that a given predictor does not increase sufficiently the accuracy of estimating the criterion, he should drop that predictor from his equation. For example, a given predictor may not add much to the decision to admit candidates to a teacher-training program, but it might be helpful in placing candidates within the parts of the program; this implies that this test should not be required for admission to the program but only for admitted candidates.

- The regression equation linking the set of predictors to the criterion will be given for the first predictor, the first predictor plus the second predictor, and so forth, so that the test user can weight the predictors appropriately in calculating the estimated criterion score for each of the candidates.
- Computational aids will be reported so that the test user does not have to substitute predictor scores into the more complicated equation to estimate a candidate's criterion score. Using these computational aid tables, the test user can estimate the criterion score for any candidate by the simple addition of several specified values. These tabled values are produced by multiplying the weights of each predictor in the regression equation by the midpoint of each score interval on the scale of the predictor. The test user may request three computational aid tables per group (or subgroup) by indicating the predictor-criterion sets he is most interested in for this purpose.

- Expectancy tables will be supplied that report the probability that a candidate with a predicted criterion score will obtain a criterion score at least as high as the value specified by the test user. This probability will be expressed in terms of the number of chances in 100 that a candidate will reach at least this specified criterion value. Expectancy tables will be reported for any computational aids requested by the test user.
- A correlation matrix will summarize the correlations between predictor measures, and between predictor and criterion measures, in tabular form. This information can help the test user to plan future validity studies.
- Experience tables will summarize the relationship between the predicted criterion score and the obtained criterion score for those predictor-criterion sets for which the test user has requested computational aids.

IV: DATA COLLECTION

In order for a school district, teacher-training institution, or state certification office to use the NTE Validity Study Service offered by Educational Testing Service, the data must be supplied to ETS by the test user in a format compatible with the computer programs written for this purpose. Further, since there are many ways in which data can be combined in a validity study, the test user must specify exactly how the data should be analyzed. The Roster Cover Sheet and Roster Sheet described in this section will enable the test user to communicate his data and purposes to ETS in terms of predictors, criteria, and design of the validity studies. Both should be filled out by the test user only after careful thought has been given to the choice of predictors, to the choice of criteria, to the design of the validity studies, and to the choice of the samples of candidates to be studied.

The Roster Cover Sheet

The Roster Cover Sheet identifies the predictors, criteria, and design of the data that are useful to the test user. Except when comparing males and females, it is necessary to fill out a separate Roster Cover Sheet for each group of candidates, or for each combination of groups of candidates, that the test user wants to study.

A sample of a Roster Cover Sheet is shown on page 22. The numbered items below refer to its various parts.

NATIONAL TEACHER EXAMINATIONS: VALIDITY STUDY SERVICE

ROSTER COVER SHEET

ITEM 1. SHEET _____ OF _____ SHEETS

ITEM 2. CODE _____ ITEM 3. NAME OF INSTITUTION AND GROUP _____ ITEM 4. LETTER I.D. OF GROUPS ☐ ☐ ☐ ☐ ☐ ☐

ITEM 5. YEARS STUDIED _____ ITEM 6. ANALYZE DATA FOR: ☐ MEN ☐ WOMEN ☐ MEN AND WOMEN COMBINED ITEM 7. RANK IN CLASS IS ☐ ACTUAL RANK ☐ CONVERTED RANK

ITEM 8. ROSTER SHEET COLUMN NUMBER	ITEM 9. ROSTER NAME	ITEM 10. DATA PROCESSING NUMBER	ITEM 11. NAME OF VARIABLE	ITEM 12. FOR ETS USE ONLY STO NAME	ITEM 13. RANGE OF DATA HIGH	LOW
4, 5, 6	RANK IN CLASS	04	RANK IN CLASS	RANK	9999	1
12	MAIN CRITERION	08	_____	_____	_____	
13	NTE WEIGHTED COMMON EXAM TOTAL (WCET)	09	NTE - WCET	NTE - WT	990	290
14	*NTE TEACHING AREA EXAM (TAE)	10	NTE - TAE	TAE	990	250
15	OTHER PREDICTOR # 1	11	_____	_____	_____	
16	OTHER PREDICTOR # 2	12	_____	_____	_____	
17	OTHER PREDICTOR # 3	13	_____	_____	_____	
18	OTHER CRITERION # 1	14	_____	_____	_____	
19	OTHER CRITERION # 2	15	_____	_____	_____	

ITEM 14. PREDICTOR - CRITERION SETS (USE DATA PROCESSING NUMBERS GIVEN IN ITEM 10)

CRITERION	PREDICTORS
SET 1 ☐	☐ ☐ ☐ ☐ ☐ ☐
SET 2 ☐	☐ ☐ ☐ ☐ ☐ ☐
SET 3 ☐	☐ ☐ ☐ ☐ ☐ ☐
SET 4 ☐	☐ ☐ ☐ ☐ ☐ ☐

ITEM 15. COMPUTATIONAL AIDS

1. SET # ☐ FIRST ☐ PREDICTORS
2. SET # ☐ FIRST ☐ PREDICTORS
3. SET # ☐ FIRST ☐ PREDICTORS

ITEM 16. EXPECTANCY TABLE VALUES

_____ _____ _____ _____ _____ _____ _____
_____ _____ _____ _____ _____ _____ _____
_____ _____ _____ _____ _____ _____ _____

ITEM 17. CROSS - VALIDATION OF OLD EQUATION (INSERT ALL PLUS AND MINUS SIGNS)

1. _____ WEIGHT ___ OP NO. 2. _____ WEIGHT ___ DP NO. 3. _____ WEIGHT ___ DP NO. 4. _____ WEIGHT ___ OP NO. 5. _____ WEIGHT ___ OP NO. 6. _____ WEIGHT ___ DP NO. 7. _____ CONSTANT

ITEM 18. LIST THE IDENTIFICATION (I.D.) OF THE GROUPS A-F FOR WHOM DATA ARE SUPPLIED ON THE ROSTER SHEETS.

LETTER	IDENTIFICATION OF GROUP
A	_____
B	_____
C	_____
D	_____
E	_____
F	_____

* NEITHER THE NTE-TAE NOR THE NTE COMPOSITE SCORE CAN BE USED IN THE STUDY UNLESS EVERYONE OF THE CANDIDATES HAS TAKEN THE TAE IN THE SAME AREA OF SPECIALIZATION AND SINCE FEBRUARY 1964. THE USE OF THE NTE COMPOSITE SCORE FURTHER REQUIRES THAT 85% OF THE CANDIDATES HAVE BOTH WCET AND TAE SCORES AVAILABLE.

Item 1. Sheet _____ of ___ Sheets

The test user must number his Roster Cover Sheets in sequential order (1, 2, 3, etc.) with each sheet coded according to its number and the number of total sheets. For example, the third sheet of a set of four would be coded as:

Item 1. Sheet 3 of 4 Sheets

Item 2. Code

The test user should leave this item blank. This code number will be assigned by Educational Testing Service.

Item 3. Name of Institution and Group; Item 4. Letter I.D. of Groups

Sixty-five spaces are allowed for giving the name of the institution requesting the validity study and the group(s) to be studied. If Aquarius Teachers College wanted to study its seniors in elementary education, it would fill out items 3, 4, and 18 as follows:

Item 3. Name of Institution and Group Aquarius Elem. Ed. Seniors

Item 4. Letter I.D. of Groups [A] [ ] [ ] [ ] [ ] [ ]

Item 18. List the identification (I.D.) of the groups A-F for whom data are supplied on the Roster Sheets

Letter	Identification of Group
A	Elem. Ed. Senior
B	Secondary Ed. Seniors in Social Studies
C	
D	
E	
F	

If this hypothetical institution wanted to study its seniors majoring in social studies in secondary education, it would need to make out a separate Roster Cover Sheet as follows:

Item 3. Name of Institution and Group Aquarius Secondary Ed. Seniors in Soc Studies

Item 4. Letter I.D. of groups [B] [ ] [ ] [ ] [ ] [ ]

Item 5. Years Studied

The year or years during which the NTE were taken by the candidates should be entered in this space. For example, if the data include NTE scores from 1964 through 1970, this would be recorded as:

Item 5. Years Studied 64-70.

If the NTE were taken by all candidates during the same year, then only the last two digits of this one year should be recorded (for example, 64).

Item 6. Analyze data for: [ ] Men [ ] Women [ ] Men and Women combined

Separate Roster Cover Sheets do not need to be prepared in order to compare males and females separately.

If an institution wants its data analyzed separately for men and women, it should code item 6 as follows:

Analyze data for: [X] Men [X] Women [ ] Men and Women combined.

If an institution wants the data analyzed for men and women combined into a single group of candidates, it should indicate this as follows:

Analyze data for: [ ] Men [ ] Women [X] Men and Women combined.

If the institution wants the data analyzed for men separately, for women

separately, and for men and women combined into a single group, it should mark all three boxes in order to receive these three separate analyses.

Item 7. Rank in class is [ ] actual rank [ ] converted rank

If rank in class is used as a predictor, it can be reported in two ways: 1) rank in class and size of class, and 2) a converted score range of 20-80 representing rank in class.

If actual rank in class and size of class are used, the correct coding is:

Item 7. Rank in class is [X] actual rank [ ] converted rank

If rank in class is expressed in terms of a converted score that has been transformed to a 20-80 range, the correct coding is:

Item 7. Rank in class is [ ] actual rank [X] converted rank

Item 8. Roster Sheet Column Number

The numbers in the column defined by item 8 refer to the column numbers on the Roster Sheet. This is done so that the data on each candidate can be coded by the test user in the proper columns of the Roster Sheet.

Item 9. Roster Name

The Roster Name column lists the predictors and criteria of interest to the test user for the group(s) designated in Item 3 and Item 4. These variables must be coded on the Roster Sheet for each candidate.

The test user should remember that neither the Teaching Area Examinations scores nor the NTE Composite scores should be included in the data analysis

for any group unless all candidates in the group took the same TAE after February 1964. For example, it would be an error to include candidates who took the Mathematics TAE and those who took the Social Studies TAE in the same analysis for a group when using TAE Scores as a predictor or criterion.

Item 10. Data Processing Number

This column gives the numbers (04, 08--15) that identify the predictors and criteria for the computer program. The numbers in this column must be recorded in the boxes under Item 14 of the Roster Cover Sheet to indicate the predictor-criterion sets of interest to the test user.

Item 11. Name of Variable

The test user writes in this column the names of the additional predictors and criteria in which he is interested. The name of the rank in class and the NTE variables are preprinted on the Roster Cover Sheet; these variables have data processing numbers 04, 09, 10.

The test user should write the name of the criteria and additional predictors in this column across from the matching data processing number. For example, the name of the criterion in which the institution is most interested should be written in this column across from data processing number 08; the names of the additional predictors should be written across from data processing numbers 11, 12, 13; the names of the additional criteria should be written across from data processing numbers 14 and 15.

Item 12. For ETS use only, STD Name

This section should be left blank by the test user. A standard abbreviation will be supplied by ETS for each variable in the study. These standard abbreviations will be included in the report of the validity study.

Item 13. Range of Data

The institution should indicate the theoretical highest and lowest score for each variable in the validity study. The highest and lowest of the NTE scores are preprinted on the Roster Cover Sheet; for example, the highest and lowest score for the WCET are 990 and 290, respectively. If a grade-point average were used by the college as either a predictor or a criterion, the range of scores would be 4.00 to 0.00 (if a four-point grade scale were used). If some institutions describe an A average as 3.00 while others use 4.00, the two ranges cannot be included in the same validity study.

Note: The decimal point should be added to every variable which includes decimals in the test score.

Item 14. Predictor-criterion Sets

The institution must specify the predictor-criterion sets in which it is most interested. For each group studied, the institution can select up to three criteria and up to six different predictors for each criterion. The criterion should be coded in Item 14 by writing in the box under "Criterion" the data processing number from Item 10 that corresponds to the criterion (08, 14, 15). The predictors should be coded in the boxes under "Predictors" in Item 14 by writing the data processing number from Item 10 that corresponds to the predictor (04, 09--13). For example, if the institution wanted to predict the main criterion (data processing number 08) by means of the WCET score and the TAE score for seniors who take the Elementary Education TAE, this would be coded in this way:

Item 14. Predictor-Criterion Sets

	Criterion	Predictors				
Set 1.	08	09	10			

Up to four different predictor-criterion sets can be specified by the institution for each validity study. Since the multiple correlations generated by the computer program will be produced in the order specified by the test user, it is important to be careful in assigning the order of the predictors. A good rule of thumb is to arrange the predictors in a predetermined order based on some hypothesis about their importance or on the basis of their availability to the test user.

The test user should not include a composite score that is the sum of two weighted scores (such as WCET+TAE) that are in the same predictor-criterion set.

Item 15. Computational Aids

A computational aid is a simplified procedure for obtaining a predicted score on the criterion. An institution can request up to three computational aids for each group studied in the validity studies. A computational aid can be obtained by writing in Item 15 the set number (1, 2, 3, 4) of interest to the institution in the box after "Set #" and the predictors of interest in the box after "First." For example, if the institution wanted to receive a computational aid for the previous example illustrated above in Item 14, it would code:

Item 15. Computational Aids

1. Set # [1] First [2] predictors

Item 16. Expectancy Table Values

An expectancy table provides a probability of obtaining at least a certain score on the criterion. An expectancy table will be provided for each computational aid requested by the institution. The institution can specify up to seven key values of the criterion for each expectancy table. For example, if the institution mentioned in Item 15 had coded the main criterion into five levels (1, 2, 3, 4, 5), it could obtain an expectancy table for these five levels by coding the sheet as follows:

Item 15. Computational Aids	Item 16. Expectancy Table Value
1. Set # [1] First [2] Predictors	1 2 3 4 5

Item 17. Cross-validation of Old Equation

If the institution wants to cross-validate to check on the accuracy of an old equation to predict the main criterion score, it should write "Predicted Average" on the Roster Cover Sheet under Item 11 opposite data processing number 13 as follows:

Item 9	Item 10	Item 11
Other predictor # 3	13	Predicted Average

If the institution furnishes the predicted main criterion scores by coding these scores in column 17 of the Roster Sheet, it can obtain the cross-validation study merely by requesting the Set 08 predicted by the other predictor # 3 under one of the four sets of Item 14; an example of this request is given below in Set 2:

Item 14. Predictor-Criterion Sets

	Criterion	Predictors		
Set 1.	14	11	09	10
Set 2.	08	13		

If an institution does not want to go to the trouble of calculating the predicted main criterion score for the group being studied, it can furnish the old equation in Item 17 of the Roster Cover Sheet as long as the new validity study includes exactly the same predictors and the same criterion as the old study. In order to use an old equation for cross-validation, the institution must indicate the set (08) predicted by (13) as indicated above, and must insert the weights of the old equation in combination with the data processing numbers to which these weights apply; this equation must also include a constant that makes the mean of the predicted score equal to the mean of the obtained scores, and all weights must be expressed as either plus or minus. For example, if the old equation for predicting the main criterion score for candidates who had all taken the Mathematics TAE since February 1964 were: Main Criterion = +.0235 WCET + .0012 TAE (Mathematics) - .3154, and the institution wanted to cross-validate this equation, the data would be recorded as follows:

Item 17. Cross-validation of old equation

1.	2.	3.	4.
+.0235 weight	+.0012 weight	____ weight	____ weight
09 DP no.	10 DP no.	__ DP no.	__ DP no.

5.	6.	7.
____ weight	____ weight	-.3154 constant
__ DP no.	__ DP no.	

Item 18. Group I.D.

Item 18 of the Roster Cover Sheet is designed so that the institution can identify up to six different groups (or subgroups) for data analysis. It is important that the group (or subgroups) be identified on the first Roster Cover Sheet by the letters A-F to indicate the groups to which each candidate belongs, as shown in the example below. The groups are coded in columns 20-25 of the Roster Sheets.

The groups A-F are also included in Item 3 of the Roster Cover Sheet, and the I.D. of the group is defined as part of the Item 18. The example which was given for Item 3 of this chapter is repeated below:

Item 18. List the identification (I.D.) of the groups A-F for whom data are supplied on the Roster Sheets.

Letter	Identification of Group
A	Elem. Ed. Seniors
B	Secondary Ed. Seniors in Social Studies
C	
D	
E	
F	

The Roster Sheet

Once the institution using the Validity Studies Service has indicated on the Roster Cover Sheet the predictors, criteria, and predictor-criterion sets of interest, the next step is to supply the data on individual candidates to Educational Testing Service. The institution should code these data on the Roster Sheets. A sample Roster Sheet appears on page 32.

NATIONAL TEACHER EXAMINATION: VALIDITY STUDY SERVICE

ROSTER SHEET

ITEM 1. PAGE _____ OF _____ PAGES

ITEM 2. NAME OF TAE _____ ITEM 3. DATES OF TAE _____ ITEM 4. CODE _____

ITEM 5. NAME OF INSTITUTION _____

ETS LINE NO.	ITEM 6. NAME OF CANDIDATE	ITEM 7. COLUMN NUMBERS																		
		(1) SEX	RANK			(12) MAIN CRITERION	(13) (14) NTE		(15) (16) (17) OTHER PREDICTORS			(18) (19) OTHER CRITERIA		(20) (21) (22) (23) (24) (25) GROUP MEMBERSHIP						
			ACTUAL RANK		CONVERTED RANK															
			(4) RANK IN CLASS 1-9999	(5) SIZE OF CLASS 1-9999	(6) CONVERTED RANK		NTE WCET 290-990	NTE TAE 250-990	1.	2.	3.	1.	2.	A	B	C	D	E	F	
1.																				
2.																				
3.																				
4.																				
5.																				
6.																				
7.																				
8.																				
9.																				
10.																				
11.																				
12.																				
13.																				
14.																				
15.																				
16.																				
17.																				
18.																				
19.																				
20.																				

Item 1. Page ____ of ____ pages

Each page of the Roster Sheets should include the number of that page and the number of total pages of roster sheets being submitted by the institution. If the institution had 25 pages of roster sheets on candidates to be analyzed, the fifth page of these roster sheets would be coded as:

Item 1. Page 5 of 25 pages

Item 2. Name of TAE

The single Teaching Area Examination (TAE) taken by all of the candidates who are listed on each roster sheet should be entered in this item. If the test user wanted to use the Social Studies TAE as a predictor, for example, item 2 would be coded as:

Item 2. Name of TAE Social Studies

If the TAE is not being used as a predictor for any of the candidates listed on a roster sheet, write the word "NONE" in Item 2 to indicate that fact.

Item 3. Dates of TAE

The years in which the NTE Teaching Area Examinations were taken by the candidates should be coded in this item. If the TAE had been taken by all social studies candidates between 1966 and 1970, this would be coded as:

Item 3. Dates of TAE 66-70

Item 4. Code

The institution should leave this item blank. A special code number will be assigned by Educational Testing Service.

Item 5. Name of Institution

The name of the institution using the Validity Study Service should be written in Item 5 of each Roster Sheet. The name given here should match the name of the institution given in Item 3 of the Roster Cover Sheet.

ETS Line Number

Each student receives this unique identification number. Please do not skip lines unless a coding error is made that cannot be changed, in which case you should cross out that line entirely.

Item 6. Name of Candidate

The data on each candidate should be coded on a separate line. The name of each candidate (last name first, first name last) should be written in this item. For example:

Item 6. Name of candidate

Jones, Charlene

Smith, Paul

Item 7. Column Numbers

Data on each candidate should be coded in the proper column. The data for columns 4, 5, 6, and 12-19 match the data specified under Item 8 of the Roster Cover Sheet. It is essential that the test specified by the institution in Item 11 on the Roster Cover Sheet be coded in the matching column on the Roster Sheets.

Column 1. Sex

The sex of each candidate must be coded in this column. Identify males by the letter "M" and females by the letter "F." If <u>all</u> candidates on the individual Roster Sheet are the same sex, enter the appropriate letter (M or F) on line 1 next to the name of the first candidate on the sheet and draw a line down through the remaining spaces in column 1 to indicate that all candidates are the same sex.

Column 4. Rank in Class

If the institution is using rank in class as a predictor, the rank of each candidate in his class should be entered in this column. The rank can range from 1 to 9999, and if the rank includes a decimal for tied scores, the decimal point and subsequent numbers following the decimal point should be dropped.

Column 5. Size of Class

If the institution is using rank in class as a predictor, the size of the class in which the rank occurred should be entered in this column. The size of the class can range from 1 to 9999. For example, if the candidate was ranked 46th in a class of 181, columns 4 and 5 would be coded as:

Actual Rank

(4)	(5)
Rank in class	Size of class
1-9999	1-9999
46	181

Note: If ranks in class are coded as percentiles, these should be reported as ranks in a class of 100; for example, a candidate who ranked at the 85th percentile should be reported as ranking 15 in a class of 100. Note that rank in class in the case of percentiles should be converted by the formula:

Actual Rank = 100 - (Percentile Rank)

Example: If the percentile rank is 85,

Actual Rank = 100 - 85 = 15

Note: If ranks in class are coded as top third, top quarter, and so forth, these should be reported as ranking 1 in 3 and 1 in 4, respectively. A candidate ranking in the third quarter should be reported as ranking 3 in a class of 4.

Column 6. Converted Rank

If the institution is using rank in class as a predictor, and if the ranks in class are already computed in converted-score form that has a range of 20-80 for all the scores in the sample, then this converted rank should be coded in column 6. Only those ranks that are already converted to a 20-80 score range should be coded in column 6.

Column 12. Main Criterion

The score of each candidate on the main criterion should be coded in this column. As many as three digits, plus the decimal point, may be used for this criterion; if the decimal point is omitted, ETS will assume that it belongs immediately to the right of the last reported digit. If course grades are used as a criterion, they must be converted from letter grades to a numerical scale (e.g. 0.00-4.00). All scores on the main criterion must be coded on the same scale.

Column 13. NTE WCET

The NTE Weighted Common Examination Total (WCET) score for the candidates can be entered in this column if the institution wants to use the WCET as a predictor in the study.

Column 14. NTE TAE

The NTE Teaching Area Examination score for the candidates can be entered in this column if the institution wants to use the TAE as a predictor for the group of candidates being studied. TAE scores can be used as predictors only if all of the candidates being studied took the TAE in the same area of specialization and after February 1964.

Columns 15-17. Other Predictors 1, 2, 3

If the institution wants to use predictors other than the ones already preprinted on the Roster Cover Sheet, it can code the scores for up to three additional predictors in columns 15-17; scores on the same predictor measure must be coded in the identical column. For each predictor used, all the data must be reported on the same scale. As many as three digits, plus the decimal point, may be used for each predictor. If the decimal point is omitted, it will be assumed that it belongs immediately to the right of the last reported digit.

If a predicted score on the main criterion is to be used in order to cross-validate an old equation, this score must be entered in column 17 ("Other Predictor 3"); this variable should then be identified as "Predicted Average" under item 11 of the Roster Cover Sheet next to data processing number 13.

Columns 18, 19. Other Criteria 1, 2

Two additional criterion scores other than the main criterion can be coded by the institution in these columns. Scores on the same criterion measure must be coded in the same column. For each criterion used, all data must be reported on the same scale. As many as three digits, plus the decimal point,

may be used for each criterion. If the decimal point is omitted, it will be assumed that it belongs immediately to the right of the last reported digit. If letter grades are used, they must be converted to a numerical scale (e.g. 0.00-4.00).

Columns 20-25. Group Membership

In columns 20, 21, 22, 23, 24, 25 of the Roster Sheet, the institution should identify the group membership of each candidate in the study. Each candidate can be coded as a member of up to six groups by coding the letter of the appropriate group underneath the column set aside for that letter. For example, if the candidate were a member of group "D," the letter "D" would be written under column 23 for that candidate.

The groups A-F should already have been specified in Item 18 on the first Roster Cover Sheet, and columns 20-25 of the Roster Sheet are set up to match these group code designations. The title of each group should have been coded as part of Item 3 on one of the Roster Cover Sheets.

Columns 20-25 of the Roster Sheet should not be used to classify candidates by sex. Column 1 of the Roster Sheet is for this purpose.

REFERENCES

1. Cooperative Tests and Services. Constructing and using local norms. Princeton, New Jersey: Educational Testing Service, 1964.

2. Cronbach, L. J. Essentials of psychological testing, third edition. New York: Harper and Row, 1970, P. 127, P. 406.

3. Cronbach, L. J. Test validation. Educational measurement, second edition, edited by Robert L. Thorndike. Washington, D. C.: American Council on Education, 1971, Pp. 443-507.

4. Cronbach, L. J. and Quirk, T. J. Test validity. The encyclopedia of education. New York: Macmillan, 1971, 9, Pp. 165-175.

5. Ghiselli, E. E. Theory of psychological measurement. New York: McGraw-Hill, 1964, P. 361, formula 11-8.

6. McNemar, Q. Psychological statistics, third edition. New York: John Wiley & Sons, Inc., 1962.

7. The National Teacher Examinations. The National Teacher Examinations: interpretation of scores. Princeton, New Jersey: Educational Testing Service, 1969.

8. The National Teacher Examinations. Bulletin of information for candidates 1970-71. Princeton, New Jersey: Educational Testing Service, 1970.

9. The National Teacher Examinations. Prospectus for school and college officials. Princeton, New Jersey: Educational Testing Service, 1970.

10. The National Teacher Examinations. Guidelines for using the National Teacher Examinations. Princeton, New Jersey: Educational Testing Service, 1971, P. 8.

GLOSSARY OF KEY TERMS

Base Rate: the percent of candidates who normally succeed when a particular test is not used for selection.

Common Examinations: that part of the National Teacher Examinations intended to measure background information common to all propective teachers regardless of teaching level or subject-field speciality. Consists of subtests in Professional Education and General Education. The weighted total test score on the Common Examinations is called the Weighted Common Examinations Total (WCET) score.

Composite NTE Score: the sum of the Weighted Common Examination Total score and the Teaching Area Examination Score (Comp. NTE = WCET + TAE) reported for all candidates who took both the NTE Common Examinations and a Teaching Area Examination on the same day. The Composite NTE Score should be included in a validity study only for those candidates who have taken the same TAE after February 1964. The Composite NTE score should never be used in a validity study if candidates have taken the TAE before that date.

Comp. NTE: see Composite NTE Score.

Computational Aids: the section of the Data Analysis Report that permits the test user to estimate the criterion score for each candidate with a minimum of arithmetical calculations.

Correlation Coefficient: a number that summarizes the extent to which the ranks of a group of candidates on one test are related to the ranks of the same candidates on another. The most commonly used correlation coefficients

range from -1.0 (the rank orders of the candidates on the two tests are in exactly the opposite order) to 0.0 (no relationship) to + 1.0 (the rank orders of the candidates on the two tests are in exactly the same order).

Correlation Matrix: a table summarizing the correlations between the predictors and criteria used in the validity study.

Criterion (Criteria): the outcome measure (or measures) the test user would like to estimate from a set of predictors.

Cross-validation: the statistical procedure by which the weights derived from one set of predictors are applied to a different but similar group of candidates to see how accurately the weights predict the criterion score from one group to another. When the weights are applied to a different group of candidates, they produce a predicted criterion score that can then be compared with the actual score to see how well the prediction equation works for different groups.

Data Analysis Report: the report of the validity study supplied to the test user by Educational Testing Service.

Equating (Equated): the statistical process by which scores on a current form of a test are made to have the same meaning as scores on an earlier form. The raw scores on the later form are put on the NTE scale in such a way that variations in difficulty of the test items and in the length of the test are compensated for. By comparing the performance of current

and past candidates on a sample of the same test items, the difference in their ability levels can be observed, and the mean and standard deviation of the candidates on the later form of the test can be adjusted to reflect this difference. For example, a score of 650 on the WCET in 1971 can be assumed to have essentially the same meaning as a score of 650 on the WCET in 1961 or in any other year prior to 1971.

Expectancy Table: a table that reports the probability (expressed in terms of chances in 100) that a candidate will reach at least the specified value of the criterion score.

Experience Table: a table summarizing the relationship between the predicted criterion score and the actual obtained criterion score for the group of candidates being studied.

Frequency Distribution: the number of candidates (N) who have test scores within specified score intervals ranging either from high scores to low scores or from low scores to high scores.

General Education: one of the two major subtests of the Common Examinations of the National Teacher Examinations. The General Education subtest assesses achievement in Written English Expression; Social Studies, Literature, and the Fine Arts; and Science and Mathematics. The General Education subtest scores are not equated from form to form and should not be used as either predictors or criteria in validity studies except with extreme caution.

Linear Multiple-regression Model: the statistical model that determines the correlation between two or more predictors and a criterion. This model is based on the assumption that a higher degree of ability in one area of knowledge or skill can compensate for a lower ability in another type by averaging those abilities considered to be important by the test user. The procedure by which more than one predictor is used to estimate a criterion score is called multiple regression. The multiple-regression equation states the weights that should be applied to each predictor in order to best estimate the criterion score for the sample of candidates being studied. The set of weighted predictors is added to a constant number to produce a predicted criterion score for each candidate with a mean predicted score equal to the mean achieved score.

Mean Score: the average resulting from the sum of test scores divided by the number of scores.

Multiple Regression: see Linear Multiple-regression Model.

Optional Examination: see Teaching Area Examination.

Percent Below: the percent of candidates who have test scores below the lowest score in the specified score interval.

Predictive Validity: the correlation coefficient resulting from the prediction of a criterion score that occurs at a later point in time from the scores of the group of candidates on the predictor test that occur at an earlier point in time.

Predictor: the test used to estimate scores on some outcome measure.

Predictor-criterion Sets: the particular combination of predictors and criterion scores that are of interest to the test user.

Professional Education: one of the two major subtests of the Common Examinations of the National Teacher Examinations. The Professional Education subtest assesses achievement in Psychological Foundations of Education, Societal Foundations of Education, and Teaching Principles and Practices. Scores on this subtest are not equated from form to form; thus, they should not be used as either predictors or criteria in validity studies except with extreme caution.

Rank in Class: the rank of each candidate in a class of a given size.

Regression Equation: the mathematical equation that specifies the weights that need to be applied to each predictor to result in the best prediction of the criterion score for a particular group of candidates.

Roster Cover Sheet: the standard cover sheet that describes the type of validity study requested by the test user. A separate Roster Cover Sheet must be filled out by the test user for each group of candidates being studied.

Roster Sheet: the standard coding sheet that allows the test user to specify the scores on the predictors and criteria for each candidate being studied and to identify the subgroups to which each candidate belongs.

Scatter Diagram: the "picture" resulting from the simultaneous coding of the predictor scores for a group of candidates along one dimension of a two-dimensional graph and the matching criterion score for each candidate along the other dimension of the graph.

Selection: the process by which some candidates are chosen to participate in a particular experience while others are rejected.

s.e.meas.: see Standard Error of Measurement.

Set of Predictors: the combination of predictors that are of interest to the test user.

Standard Deviation: the measure of the spread of scores about the mean score. In general, scores that are spread out over a larger range of scores have a larger standard deviation than scores spread out over a smaller range of scores.

Standard Error of Estimate: the standard deviation of the criterion scores of those candidates who have the same predictor score. The difference between the predicted criterion score and the actual criterion score is the error in the prediction for each candidate. In general, two-thirds of the candidates with a given score on the predictor will score within one standard error of estimate (plus or minus) of their predicted criterion score. As the correlation between the predictor and the criterion increases, the error in prediction decreases.

Standard Error of Measurement: an indication of the error in the test score due to the less-than-perfect reliability of the test. If the standard error of measurement of a test is 30, the chances are 2 to 1 (68 chances in 100) that a candidate's score on the test will be within 30 points (plus or minus) of what his score would be if there were no error.

Subgroup: a group of candidates included by definition in some larger group. For example, if the total group of candidates is studied separately for males and for females, the group of males and the group of females each constitute a subgroup.

Summary of Statistics: the section of the Data Analysis Report that indicates the degree of improvement of the prediction of the criterion when two or more predictors are used to estimate the criterion, the correlation of each predictor with the criterion, the mean, standard deviation, and standard error of estimate for each predictor, and the mean and standard deviation of the criterion.

TAE: see Teaching Area Examination

Teaching Area Examinations (originally called the Optional Examinations): 24 subject-area speciality examinations were offered by the National Teacher Examinations during 1971. Since the TAE were not equated from form to form until February 1964, no scores earned prior to that date should be included in any validity study. Only TAE scores in the same subject-area specialty should be included in the data analysis for any group of

candidates. It is incorrect to mix together scores on different TAE (e.g., mathematics and social studies) in the same validity study since the test scores do not have the same meaning (i.e., the TAE differ in test content and difficulty) from one TAE to another.

Variable: a property, behavior, or trait in which candidates differ among themselves.

WCET: see Weighted Common Examinations Total Score.

Weighted Common Examinations Total (WCET) Score: the weighted combination of the Professional Education subtest and the General Education subtest of the NTE Common Examinations. Since the WCET scores have been equated from form to form since 1940, it is possible to mix WCET scores in the same data analysis even though the test was taken in different years.