INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of "sectioning" the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.
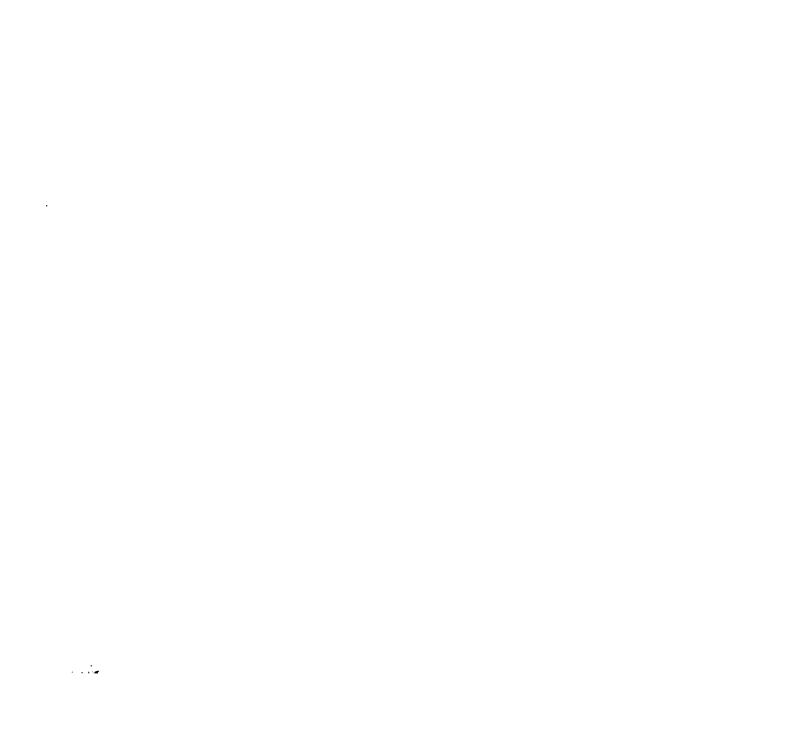
8521442

Anderson, Kenneth Edward

THE DEVELOPMENT OF AN ITEM POOL FOR ASSESSING PRACTITIONER KNOWLEDGE OF THE ATTRIBUTES OF RECOMMENDED GRADING AND MARKING SYSTEMS

The University of Nebraska - Lincoln   PH.D.   1985

University
Microfilms
International   300 N. Zeeb Road, Ann Arbor, MI 48106

THE DEVELOPMENT OF AN ITEM POOL FOR ASSESSING

PRACTITIONER KNOWLEDGE OF THE ATTRIBUTES

OF RECOMMENDED GRADING AND MARKING SYSTEMS

by

Kenneth E. Anderson

A DISSERTATION

Presented to the Faculty of

The Graduate College of the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Interdepartmental Area of

Administration, Curriculum, and Instruction

Under the Supervision of Professor Frederick C. Wendel

Lincoln, Nebraska

May, 1985

TITLE

THE DEVELOPMENT OF AN ITEM POOL FOR ASSESSING PRACTITIONER
KNOWLEDGE OF THE ATTRIBUTES OF RECOMMENDED GRADING AND
MARKING SYSTEMS

BY

Kenneth E. Anderson

APPROVED	DATE
Frederick Wendel	April 30, 1985
Ward Sybouts	April 30, 1985
Alvah Kilgore	April 30, 1985
Robert Egbert	April 30, 1985

SUPERVISORY COMMITTEE

GRADUATE COLLEGE   **UNIVERSITY OF NEBRASKA**

THE DEVELOPMENT OF AN ITEM POOL FOR ASSESSING

PRACTITIONER KNOWLEDGE OF THE ATTRIBUTES

OF RECOMMENDED GRADING AND MARKING SYSTEMS

Kenneth E. Anderson, Ph.D.

University of Nebraska-Lincoln, 1985

Advisor: Frederick C. Wendel

The purpose of this study was to create a pool of
test items to assess knowledge of principles and practices
associated with the rational, psychometric, legal, and
learning domains of grading and marking systems, for which
reliability estimates and content validity characteristics
would be established. The items were created and validated
by implementing a plan to enhance content validity and as-
sess reliability. Two research objectives were specified
to assist in validating the test items. These were:

    1. To ascertain the extent of agree-
ment among experts, with content
knowledge congruent with the four
domains of grading and marking
systems, as to what information
is essential for practitioners to
either know or apply to implement
grading and marking systems that
are consistent with recommended
practices.

    2. To discover the extent of agree-
ment among teachers who regularly
apply grading and marking systems
as to what information is essen-
tial to either know or apply to
implement recommended grading and
marking systems.

Material to be included in the test items was chosen

by conducting a content analysis of research findings associated with studies of grading and marking systems. The items were validated by collecting and analyzing information from a professional jury and a sample of Nebraska teachers.

Data regarding the nature, content, and merit of the test items were collected from 4 school psychologists, 6 educational psychologists, 7 measurement and evaluation specialists, and 2 school attorneys; all had been identified as nationally known experts in their fields. A sample of 167 Nebraska teachers completed tasks to provide information for reliability estimates and item analyses.

Although content validity was built into the items and a pool of items representing information that experts agreed was essential to know or apply was developed, the items in the pool possessed little reliability. Suggestions for modifying the items to enhance their reliability were outlined, and a list of recommendations for developing and validating grading and marking systems was delineated.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

Traditionally courts have granted broad discretionary
power to individual school districts to establish academic
grading standards (Gatti & Gatti, 1975; McCarthy & Gambron,
1981). As districts have exercised this power to devise
grading and marking plans responsive to their diverse
needs, a plethora of evaluative systems have prospered.
Systems which have evolved have not always incorporated
sound rational, psychometric, legal, and learning premises
into their methods, and numerous problems have been asso-
ciated with their application to document pupils' academic
progress (Stallings, 1968; Hedges, 1969; Stanley & Hopkins,
1972; Hills, 1976; Broadfoot, 1979; Branwaite, 1981; Stef-
anelli, 1981; The National Commission on Excellence in Ed-
ucation, 1983).

From the 1950's through the mid-1980's, definitional
inconsistencies have plagued the terms, grading and mark-
ing, and researchers, as well as practitioners, have demon-
strated confusion over related terminology. Grading, and
most likely marking, has been described as one of the most
persistent and vexing problems in American education (Sla-
vin, 1980).

Despite the breadth of problems associated with grad-
ing and marking practices, little has been done to resolve,
or even to clarify, controversies associated with these

problems. Grading and marking systems have been character-
ized as necessary components of educational programs (Ebel,
1965; Dressel, 1976; Hills, 1976; Miles, 1977; Karmel,
1978; Ebel, 1979; Bloom & Bourdon, 1980; Bippus, 1981).
Researchers have failed to agree on the merits and liabil-
ities of any single system.

Grading and marking systems are envisioned as the in-
teraction of attributes of four domains (See Figure 1).
Rational, psychometric, legal, and learning domains inter-
act to comprise a single grading and marking system. The
quality of assumptions about, knowledge of, and applica-
tions for the four domains determine the legitimacy of the
system. Inadequacies inherent in attributes or applications
within any of the four domains can invalidate the entire
system's worth.

Persons developing and implementing a grading and
marking system should be knowledgeable about the assump-
tions, attributes, and practices inherent in each of the
four domains that comprise the system. Failure to properly
acknowledge principles, or to accommodate for inadequacies
of practice within any domain, is likely to result in the
implementation of a system that is weak in either reliabil-
ity, validity, or both.

GRADING
(OR MARKING)
SYSTEM

RATIONAL   PSYCHOMETRIC

INPUTS

LEGAL   LEARNING

Figure 1

The Four Domains of Grading and Marking Systems

Adapted From Edgar A. Kelley (1983)

PURPOSE OF THE STUDY

The purpose of this study was to create a pool of test items to assess knowledge of principles and practices associated with the rational, psychometric, legal, and learning attributes of grading and marking systems, for which reliability estimates and content validity characteristics would be established.

RESEARCH OBJECTIVES

Two research objectives were specified to assist in validating the test items. These were:

1. To ascertain the extent of agreement among experts, with content knowledge congruent with the four domains of grading and marking systems, as to what information is essential for practitioners to either know or apply to implement grading and marking systems that are consistent with recommended practices.

2. To discover the extent of agreement among teachers who regularly apply grading and marking systems as to what information is essential to either know or apply to implement recommended grading and marking systems.

SIGNIFICANCE OF THE STUDY

In spite of the prevalence of grading and marking systems in American educational institutions, widespread and vexing problems have been associated with their use,

suggesting that practices often utilized in schools may be inappropriate (Stanley & Hopkins, 1972; Slavin, 1980). Practitioners should be familiar with attributes associated with the rational, psychometric, legal, and learning domains that comprise grading and marking systems if they are to design and implement systems that are consistent with recommended practices.

The extent of educators' familiarity with attributes of these four domains is not known. In this study, a pool of test items was developed as an initial step toward constructing a test instrument to diagnose educators' levels of knowledge across the four domains of grading and marking systems. Once a finalized test instrument has been completed and all items validated, it should be administered to practitioners as a diagnostic tool.

If practitioners score well on a test instrument which is comprised of items that represent information about the principles and practices inherent in the four domains of recommended grading and marking systems, then they probably possess at least the threshold levels of expertise necessary to apply recommended grading and marking systems in an appropriate manner. If practitioners do not score well on the test instrument, knowledge deficits inhibiting the appropriate application of principles and practices inherent in the four domains of recommended grading and marking systems may be pinpointed. Learning activities corresponding

to areas of diagnosed weaknesses may be designed, and pre-service training, inservice training, or both, can be implemented to ameliorate the identified deficits.

DEFINITIONS

Definitions for substantive terms are as follows:

Scoring - The process of correcting student work and calculating a raw or derived score based on the assessed performance.

Grading - The attachment of a grade (usually a letter or number symbol) to a student product based upon comparison of the performance score to some standard.

Marking - The assignment of a composite grade symbol (usually a letter or number) to represent the cumulative achievement of a student as measured on a variety of products.

Grading system - A set of methods for logically deriving grades in a specific, orderly manner.

Marking system - A set of methods for logically deriving composite marks in a specific, orderly manner.

Assessment - The assignment of numbers to objects according to logical rules.

Evaluation - To place values based on information.

Appraisal - Evaluation based on numeric values assigned in assessment.

Teachers - Persons listed in the Nebraska Educational Directory as teachers.

Rational domain - Elements of the reasons for the existence of grading and marking systems, as well as the present status of their application.

Psychometric domain - The measurement capacities of grading and marking systems.

Legal domain - Legal concerns in assigning grades and marks.

Learning domain - The aspects of grading and marking systems that, as defined by contemporary learning theory, serve to reward and punish students.

LIMITATIONS OF THE STUDY

Conclusions and implications about the research objectives of the study are limited to professionals who possess knowledge congruent with the four domains of grading and marking systems (school psychologists, educational psychologists, measurement and evaluation experts, and school attorneys) and Nebraska teachers.

ASSUMPTIONS OF THE STUDY

Assumptions accepted as the basis for interpreting the results of the study were:

1. Grading and marking systems are comprised of attributes of four domains: rational, psychometric, legal, and learning.

2. School psychologists, educational psychologists, measurement and evaluation experts, and school attorneys possess expertise congruent with attributes of the rational, learning, psychometric, and legal domains of grading and marking systems, respectively.

3. Multiple-choice items were considered to be the most efficient means for collecting data on knowledge of the attributes of grading and marking systems, as they have demonstrated the capability to sample a wide variety of data sources, are easily administered and scored, and yield information that can readily be subjected to statistical scrutiny (Nunnally, 1967).

METHODOLOGY

The pool of test items produced in this study was completed by implementing a developmental plan to establish item reliability and content validity. Specific details of that plan are outlined in Chapter III. Material for inclusion in the test items was determined by conducting a content analysis of research findings reported in Chapter II. The test items were created to reflect a multiple-choice format. The items were refined by three university professors. Respondents from a jury comprised of 40 professionals from the fields of school psychology, educational psychology, evaluation and measurement, and school law reviewed the items and provided information to assist with item validation. Finally, respondents from a sample of 250 Nebraska teachers answered the test items to provide information to complete item analyses and reliability estimates.

A final pool of test items was chosen by selecting the items with the best psychometric and validity characteristics. Information to assess two research objectives was gathered as the test items were validated by the professional jury and the Nebraska teachers.

CHAPTER II

REVIEW OF LITERATURE

<u>INTRODUCTION</u>

Research findings reported from the 1950's through
the early 1980's have indicated that grading systems, and
probably marking systems as well, are invalid measures of
knowledge, destructive to the development of student self-
esteem, motivationally misdirecting, and anxiety-producing
(Anderson, 1982). There is also research to suggest that
marking systems provide useful information, have a positive
influence on pupil behavior, and accurately predict stu-
dents' future academic achievement (Hull, 1980). Confusion
concerning grading and marking practices seems to be a nat-
ural outcome of this research.

This chapter will appraise research findings associated
with the rational, psychometric, legal, and learning domains
of grading and marking systems. The chapter has been devel-
oped in three sections. First, operational definitions for
substantive terms and processes are proposed. Second, ra-
tional, psychometric, legal, and learning domains are delin-
eated in separate sections. Last, a summary of pertinent
research findings is presented.

OPERATIONAL DEFINITIONS

Educational literature is replete with information on grading and marking practices. Unfortunately, there appears to be conceptual confusion among theorists and researchers regarding the concise meaning of substantive terms. Oftentimes the terms scoring, grading, and marking are used interchangeably, when in fact, they do not connote synonomous processes.

Operational definitions for scoring, grading, and marking are presented in order to neutralize potential conceptual confusion (Green, 1970). Green (1970) has developed a useful categorization scheme for these terms and their processes, and his thoughts will be reflected in the ensuing discussion.

Scoring refers to the process of correcting student work and calculating a raw or derived score based on the assessed performance. The process of scoring may be viewed as the initial step in measuring and documenting student achievement, as it generates data that will be coded into a grade.

Grading involves the attachment of a grade (usually a letter or number symbol) to a student product based upon comparison of the performance score to some standard. In the process of grading, scores (raw or derived) are transformed into a grade symbol that represents a level of

achievement. Grading sets a value on a score by affixing a symbol that connotes a specified level of achievement (Nunnally & Ator, 1972; Lien & Lien, 1980). Brown (1971) has reported that the process of assigning a grade usually involves three steps. First, a raw score is obtained. Second, the raw score is changed into some type of derived score. Third, the derived score is transformed into a symbol that connotes the designated level of achievement.

Marking consists of a similar, but more generalized process than grading. Marking consists of assigning a composite grade symbol to represent the cumulative achievement of a student as measured on a variety of products. Marking is the process of assigning a composite grade symbol (mark) based upon the consideration of a series of individual grades.

There are some subtle conceptual and practical differences between the processes delimited by the terms scoring, grading, and marking. Unfortunately, as theorists and researchers have been generally insensitive to these differences, they have not clearly distinguished between conceptualizations of grading and marking practices. As a result, the categorization, interpretation, and generalization of relevant research findings has often been confused.

THE RATIONAL DOMAIN
OF GRADING AND MARKING SYSTEMS

Rationale for the Existence of Grading and Marking Systems

There is consensus among members of the educational community that there is a need for some form of overt structure to monitor, chart, and report systematically the academic progress of students. Grading and marking systems have traditionally provided the structure to fulfill this purpose. Grading (and marking) systems have been characterized as necessary components of educational programs that serve as means to communicate the status of student academic achievement (Ebel, 1965; Dressel, 1976; Hills, 1976; Miles, 1977; Karmel, 1978; Ebel, 1979; Bloom & Bourdon, 1980; Bippus, 1981). Thus, grading and marking systems are given high regard as methods to monitor, chart, and report students' levels of academic achievement (Davis, 1964; Payne, 1968; Hills, 1976; Ebel, 1979).

Other uses of grades and marks have centered around their utility in facilitating administrative judgments. Grading and marking systems have provided information to assist with decisions regarding student promotion, admission to special classes and programs, the presentation of academic honors, and the determination of student eligibility for scholastic and extracurricular activities (Mitzel, Best, & Rabinowitz, 1982).

Grading and marking systems fulfill several important

educational purposes, and herein lies the rationale for
their existence. School personnel use grading and marking
systems to monitor, chart, and report student academic
progress, and to provide information to assist school of-
ficials in making decisions about students. Each of these
uses suggests a necessary purpose for school use; collec-
tively, they imply a series of essential record keeping,
communication, and administrative purposes central to the
operation of schools.

Although educators generally agree that grading and
marking systems are necessary components of educational
programs, they adamantly disagree on the merits and lia-
bilities of specific grading and marking practices promul-
gated by individual systems. As a result, there is no sin-
gle system accepted as the "state of the art". Many
schemes for grading and marking are presently used in
schools (Brown, 1971).

Designing and implementing grading and marking sys-
tems is the responsibility of individual school districts.
At present, there is no legislation to mandate the use of
one particular system across states. As educational insti-
tutions respond to differential controls based upon the
amount of regulation exercised by each state or local mu-
nicipality, a plethora of practices is encouraged. The
number of specific systems used in schools could potential-
ly be so great as to defy generalization (Anderson, 1982).

Nonetheless, there have been attempts to identify global types of systems used in schools.

Types of Grading Systems

Terwillinger (1971) has proposed one system for categorizing types of grading systems. His scheme promotes four classifications for grading systems: 1) systems that utilize absolute standards, 2) systems that measure achievement in reference to students' measured abilities, 3) systems that measure students' achievement in reference to the achievement of peers, and 4) systems that base measurement functions on individual student growth. Other researchers have also offered schemes for classifying types of grading systems, but have used differing category definitions and dimensions (Brown, 1971; Nunnally, 1972; Smith & Adams, 1972; Hills, 1976; Gronlund, 1981). In this chapter, Terwillinger's categorization scheme will be used as a guide to outline relevant research findings.

Grading by absolute standards. Grading systems which incorporate absolute standards measure student achievement against some absolute, pre-specified standard (Terwillinger, 1971). Each individual's performance is independent from that of other students (Brown, 1971). An example of an absolute standard system is as follows. All students who score between 93 and 100% correct on an assigned task receive the highest grade (e.g., an A if a letter system is used). Students who obtain scores of 85-92% correct

receive the next highest grade symbol. Students who score
from 70-84% correct receive the next specified grade sym-
bol. Students scoring 60-69% correct receive the next low-
est symbol, and anyone scoring below 60% correct is as-
signed the lowest grade symbol.

Grading by comparing achievement to ability. Systems
which measure students' achievements in reference to their
abilities assess comparisons between expectancies based
upon an individual student's abilities and actual achieve-
ment (Terwillinger, 1971). Students who achieve beyond
the level expected of them are assigned the highest grades
while those who achieve below expectations receive lower
grades. An example of a system based on comparisons be-
tween students' abilities and achievements is as follows.
First, measures of ability are obtained during the first
week of class and placed on file, and then expectancies
are determined. Ability scores of students are ordered
from high to low and divided into three equal groups to re-
flect high, medium, and low ability, respectively. After
achievement tests are given, the average achievement score
is calculated within each ability level grouping, and the
average score obtained in this manner becomes the expected
achievement score of students within each category. Final-
ly, scores reflecting discrepancies between ability and ex-
pected achievement are derived for each student. Student
scores that have the greatest positive discrepancies in

favor of achievement are given the highest grades, and those with the highest negative discrepancies are given the lowest grades. Other grade symbols are then assigned for the remaining categories.

Grading by comparing achievement to that of peers. Grading systems which are premised on measuring students' achievement in reference to that of their peers assume the achievement levels of other students as the preferred frame of reference (Lindvall, 1961). The achievement of a student's norm group provides a distribution of scores against which the performance of an individual is compared. Such comparisons typically take the form of percentile ranks or standard score conversions. A grading system based on individual comparisons to the achievement of peers (by the use of percentile ranks) includes the following. The scores obtained by class members are rank ordered and percentile ranks are computed for each. Any scores that attain the 91st percentile rank or higher are given A's (assuming the use of a letter grading system). Scores that range from the 66th to the 90th percentile ranks are awarded B's. Scores that range between the 21st to the 65th percentile rank are given C's. D's are assigned to scores ranging between the 6th and the 20th percentile ranks, and F's are assigned to scores falling below the 5th percentile rank. The exact cutoff points for determining the grades may be set according to absolute or relative standards,

depending on the preference of the teacher.

Grading by individual student growth. Grading systems based on student growth can be considered special types of systems which measure student achievement in reference to the achievement of peers. A grading system which is based on measuring the growth of an individual student uses the degree of improvement a student exhibits as the criterion for determining levels of achievement (Terwillinger, 1971). Measurements are performed at two (or more) points and results are compared. Achievement is assessed as the degree of growth students display between measures. Grades are then assigned to students by either comparing the amount of each student's growth against a pre-specified standard, or by comparing the amount of growth of individual students against that of their peers. An example of a system based on measuring student growth against that of peers is as follows. Students who demonstrate the most improvement from pretest to posttest scores measuring knowledge of a segment of course content receive the highest grades (e.g., A's if a letter grading system is used). Students who demonstrate lesser amounts of gain are awarded B's, C's, D's, or F's respectively.

Types of Marking Systems

The processes of grading and marking represent philosophically similar constructs. Marking systems can also be subsumed under the same four system types outlined for

grades. Examples of marking systems within these four
categories are congruent with the examples of grading sys-
tems which have been provided, but composite marks, not
grades, are the product in each case.

Means for Expressing Grades and Marks

Letters, numbers, percentage figures, or descriptive
statements may all be used to express student achievement
within the four categories of systems described above. In-
dividual school districts have utilized varying types of
grade symbols based upon tradition or preference, and the
types of symbols used by districts exhibit considerable
variation (Anderson, 1982). Researchers have offered data
to specify rates of usage of the types of symbols utilized
in grading and marking systems.

Hedges (1969) and Newman and Wise (1975) have suggest-
ed that the most popular technique for expressing grades
(and marks) in elementary schools has been the use of let-
ter grade symbols. Letter symbols have also been used fre-
quently to report the achievement of high school students.
Results of a national survey of over 600 schools conducted
by the National Education Association in 1966 indicated
that letter symbols or numeric symbols were used to report
grades in almost 80% of the schools sampled (Stanley & Hop-
kins, 1972). A second national survey conducted by the Na-
tional Education Association in 1971 indicated that letter

symbols were assessed to be the most popular means for expressing student achievement. Descriptive word grades and narrative reports were also found to be popular, but to a lesser degree than letter symbols (Newman & Wise, 1975). Data gathered from 1069 high schools as part of the National Longitudinal Study of the High School Class of 1972 indicated that 67% of the surveyed schools used letter symbols only to report student progress; 16% used percentages; 5% used a combination of letter and percentage symbols; and other types, combinations of types, and missing data accounted for the remaining 11% of the schools (Robinson, 1978).

Hills (1976) reviewed secondary school grading (and marking) systems and reported that the most commonly used system in 1976 (68% of the schools surveyed) was the utilization of the letter symbols A through F to document student performance, with A representing the highest grade (or mark) and F the lowest. These letter symbols were assigned to student products based on individual teacher standards. The second most commonly used system (16% of the schools) was the assignment of grades or marks based on a fixed numerical scale. Letter symbols were still assigned, but they were assigned based on a fixed scale using the following percentages of product items correct: 90-100=A, 80-89=B, 70-79=C, 60-69=D, below 60=F. The third most commonly used system (3% of the schools) utilized the numbers

1 through 5 to report pupil progress. Student scores on
tasks were transformed to this scale according to a pre-
specified criterion. The least commonly applied system
(less than 2% of the schools) was the Pass-Fail system in
which only two grades or marks were issued: one to indi-
cate success (P), and the other to indicate failure (F),
as based on pre-specified criteria. Hills (1976) did not
specify the number of schools surveyed, the nature of his
sample, or techniques used for collecting and reporting
data.

There have been reports of schools moving away from
traditional letter or number grading systems like those
listed above, but these changes appear short-lived. Some
schools have attempted to move away from traditional means
to grade students, but have been limited in their efforts
due to external pressure from administrators, parents,
boards of education, and teachers (Warner, 1978). In ad-
dition, some schools which have initiated non-traditional
grading practices have been forced to drop these practices
in favor of the traditional practices, thus further sug-
gesting consumer discontent with non-traditional grading
and marking practices (Hedges, 1969).

Letter symbols are probably the most widely used meth-
ods to express student grades and marks. Other forms for
expressing grades and marks have been observed to lesser
degrees, but their use has varied across school districts.

With the exception of Hills' (1976) study, the relative
usage of grading and marking systems in relation to the
four categories of standards previously discussed has not
yet been established.

Problems Associated with Grading and Marking Systems

Researchers have specified problems regarding the ap-
plication of grading systems in public K-12 schools from
the 1950's through the early 1980's. Stanley and Hopkins
(1972) have proposed a useful strategy for categorizing
grading problems and have suggested that the problems as-
sociated with assigning grades to document students' work
can be subsumed under four categories of concern:
1) grades are inaccurate and not comparable across instruc-
tors, departments, or schools, 2) grades focus upon false
and inappropriate objectives which maintain little rela-
tionship to important educational objectives, 3) grades
are limited in value as a medium of communication between
school and home, and 4) grades are responsible for a vari-
ety of detrimental side effects.

Inaccuracies of grades. Theorists have contended that
grades are inaccurate and thus, unreliable. Two research
studies have been located to support this contention.
Hedges (1969) and Branwaite (1981) have suggested that dif-
ficulties are observed comparing grades across teachers,
departments, or schools. Personality factors and personal
appearance have been postulated to influence grades to some

unknown degree. Russell and Thalman (1955) collected data on 335 seventh and eighth grade students to assess the relationship of student personality factors to grades. In this study, results of personality assessments were correlated to student grades, and a positive relationship was discovered between the grades pupils received from teachers and personality ratings of the students which were given by the teachers; the more positive the rating, the more positive the grade. Salvia (1977) divided 84 third, fourth, and fifth grade students into five categories based upon facial attractiveness, then correlated levels of attractiveness to grades obtained by the students. A significant positive correlation was reported between attractiveness and grades; the more attractive students received the higher grades, and the least attractive students received lower grades.

Increases in student grade averages have been noted recently by some researchers, and these have been attributed to: 1) permissiveness of faculty grading practices, 2) changing student populations, and 3) the ability of students to drop courses before low grades or marks are assigned (Ahmann & Glock, 1981). This inflationary syndrome would appear to contribute further to the incomparability (unreliability) of grades.

Teachers' awards of "extra credit" for additional tasks assigned in classes represent practices which may

potentially contribute to the incomparability of grades.
Differential use of these practices, as well as non-uni-
form applications by teachers, can inflate grades diverse-
ly to an unknown degree, thus exacerbating their incompar-
ability (Dressel, 1976; E. Kelley, personal communication,
March, 1983).

False focus of grades. Grades have been described as
narrow in scope and subjective in nature (Rinne, 1975).
Stefanelli (1981) has alleged that grades focus on false
and inappropriate objectives which maintain little relation-
ship to important educational goals, and that grades possess
little utility for documenting learning in important non-
academic areas such as behavior, attitude, and self-respect.

Mitzel, Best, and Rabinowitz (1982) have reported that
reliance upon grades as motivators for some students may in
fact inhibit the achievement of a major educational goal,
the development of lifelong study skills. Study skills may
extinguish when grades and marks are no longer given as mo-
tivators after graduation. Further, grades may not lead
toward increased motivation of all students (Cullen, et.
al., 1975; Yarborough & Johnson, 1980).

Cullen et. al. (1975) sampled 233 high school freshmen
and administered one of three conditions to each student.
Students were offered points (ranging from 2 to 12) for
completing an assignment, were threatened with a loss of
points (ranging from 1 to 7) if the assignment was not

completed, or were told only to complete the assignments with no consequences discussed. The authors concluded that, for some students, grades used as a negative incentive were more powerful in increasing task completion than grades used as a positive incentive. These conclusions did not suggest congruent effects across all students.

Yarborough and Johnson (1980) matched 51 seventh graders from elementary (K-6) schools which did not give grades to three control groups from 21 elementary schools which gave grades to students and assessed the students' attitudes toward school. Findings indicated the following: 1) brighter pupils from graded schools tended to have more positive attitudes toward school than brighter pupils from nongraded schools, and 2) slower pupils from nongraded elementary schools tended to have more positive attitudes toward school than slower students from graded schools. The authors concluded that grades (marks) may mean different things to different students, thus questioning the consistent motivational value of grades for all students.

Limited value of grades as means for communication. Grades have been suggested to be of limited value as a medium of communication between home and school. Hills (1976) has indicated that teachers do not effectively communicate through grades, and Stanley and Hopkins (1972) have also postulated that grades, in and of themselves, do not foster good communication between home and school.

Detrimental side effects of grades. Stanley and Hop-
kins (1972) have reported that grades are responsible for
producing a variety of detrimental side effects such as
anxiety, dishonesty, hostility, and poor mental health;
they produce negative attitudes resulting from chronic
failure, and encourage undesirable value patterns. Stal-
lings (1980) surveyed 2300 college students regarding their
perceptions of grades. Seventy-eight percent of the re-
spondents agreed that an emphasis on grades encouraged
cheating. Broadfoot (1979) has suggested that grading
alienates students and does not provide for increased
achievement motivation. Marshall (1981) has also reported
that grading alienates students, and that grading creates
a disposition toward increased dependency upon adults.
Mitzel, Best, and Rabinowitz (1982) have suggested that
students who assume they are forced to perform academic be-
haviors under penalty of earning a low grade or mark may,
as a result, develop negative attitudes toward academic be-
havior.

To summarize, findings related to present problems
with grades suggest that they are unreliable, narrowly
based, of limited communicative value, and likely to gener-
ate serious and deleterious student side effects. Even
though the authors cited in this section all defined their
studies as assessments of grading practices, confusion is
evident regarding the use of grades, marks, or both as the

object of study. Generalizations from the information which has been reported must be carefully applied.

No studies outlining specific problems with marks have been located, but marks, like grades, are probably subject to the four categories of problems outlined above. The same criticisms leveled at grades may likely be applied to marks as well.

THE PSYCHOMETRIC DOMAIN OF GRADING AND MARKING SYSTEMS

The Psychometric Attributes of Grades

What do grades measure? How well do they measure whatever it is that they measure? These two questions furnish the framework for the ensuing discussion on the psychometric attributes of grades.

In response to the first question, (What do grades measure?), researchers are in general agreement that grades at all levels of schooling should measure student academic achievement. This position was clearly stated as a recommendation of the National Commission on Excellence in Education (1983). Assessments of students' attitudes, efforts, abilities, personalities, behaviors, and attendance should not be reflected in grades (Davis, 1964; Lindeman, 1967; Mehrens & Lehmann, 1973; Gronlund, 1981). Nonetheless, there is some evidence to suggest that grades based on effort and ability continue to be used widely in elementary schools (Gronlund, 1981); Gay (1980) has suggested that

this constitutes an appropriate practice.

Grading systems which measure effort and ability, even if used only at the elementary school level provide interesting, but misleading information. Grades, as measures of academic achievement, become contaminated by the differential influence exerted by evaluative measures that attempt to assess variables foreign to academic achievement (Ebel, 1965). Grades, if influenced by "non-academic" factors, lose their utility, as they represent a nebulous compilation of elements. Meaning and comparability are compromised, and confusion of interpretation is a probable outcome (Gronlund, 1981).

Measures of attitude, effort, personality, behavior, and attendance are certainly important factors in speculating upon the relative successes or failures of students, but these should not be averaged with academic performance and reported as measures of academic competence (Mehrens & Lehmann, 1973). Evaluative comments relative to factors other than academic achievement may be reported concurrently with grades by the use of checklists or descriptive word statements; however, these should in no way be mistaken for academic grade reports (Davis, 1964). Grades should reflect students' academic achievement, nothing more, and nothing less (Davis, 1964; Ebel, 1965; Mehrens & Lehmann, 1973; Gronlund, 1981).

Ebel (1965) defined grading as a means for reporting

measurements of achievement in contrast to a means for reporting evaluations of student progress. Achievement can be derived and reported mathematically, based on focused empirical information. Evaluation, however, involves the complex interplay of many considerations unique to individuals that usually cannot be measured mathematically and applied uniformly to all students. If evaluative information is considered in reporting academic progress, difficulties are encountered in fitting this information into standardized, mathematically-derived grading systems. Efforts to do so decrease the applicability and consistency of the system, as evaluative information cannot be applied uniformly to all students.

How well do grades measure whatever it is that they measure? The answer to this question demands a qualitative line of reasoning which will be provided by assessing the relative reliability and validity that grades possess.

Measures of the reliability of grades are extremely difficult to accomplish, as there is no entirely satisfactory means to estimate the reliability of a set of grades statistically. Educators are not able to exercise the degree of experimental control necessary to obtain "pure" measures of reliability, as exams, assignments, and experimental treatments cannot ethically be manipulated or withheld from groups of students as measures of statistical balance. In practice, only rough approximations of

reliability coefficients can be obtained. Practitioners should remember that coefficients of reliability obtained in educational settings are not infallible, and that there is room for error (Davis, 1964).

Grades are measurements which are similar to tests; therefore, they may be subjected to reliability concerns philosophically similar to those for tests. Guilford and Fruchter (1973) suggest that there are many methods for estimating test reliability. These fall into three general categories: 1) internal consistency reliability, 2) alternate forms reliability, and 3) test-retest reliability. Estimates of the reliability of grades may be subjected to this same categorization scheme.

Internal consistency reliability coefficients measure the extent to which all components of an instrument assess the same outcome (Guilford & Fruchter, 1973). An instrument which has perfect internal consistency is comprised of items that all measure exactly the same outcome. For a grade to be internally consistent, it must be based on a student product that is made up of items which all measure the same end.

Alternate forms reliability coefficients appraise the degree of agreement between measures on two parallel forms of assessment (Guilford & Fruchter, 1973). Two comparable forms assessing the same content are developed and administered to groups of students. The agreement of students'

scores on the forms is checked statistically to assess the degree to which they measure the same outcome. The extent to which congruent content grades assigned by different teachers agree represents a type of alternate forms reliability.

Test-retest reliability coefficients measure the stability of scores over time (Guilford & Fruchter, 1973). A single measure is administered to the same population two or more times with a time interval between the administrations, and the degree of stability of individual scores is assessed. For a grade to demonstrate the standard of test-retest reliability, the student product on which the grade is based would be assigned at several points in time. The scoring system and the method for assigning the grade would be repeated each time, and the extent of agreement between the obtained grades would be assessed.

Most of the research available on the reliability of grading systems is difficult to analyze according to the aforementioned scheme, primarily as a result of substantive terminology discrepancies. As has been noted previously, authors are not always careful to distinguish between the processes of grading and marking, and the application of appropriate tests of reliability for these concepts is confused.

Information available on the reliability of grading practices seems to center around questions akin to those

addressed by alternate-forms types of reliability esti-
mates. This information appears to be theory-based. No
data-based studies have been located to assess the neces-
sary outcomes. Lesser, Davis, and Nahemar (1962), Davis
(1964), Hedges (1969), Gay (1980), and Branwaite, Trueman,
and Berrisford (1981) have all suggested that there is lit-
tle consistency between teachers in their grading prac-
tices. Therefore, if two teachers teach the same course of
academic study, a grade of "C" from the first teacher on a
student product assigned as a part of the course may not be
comparable to a grade of "C" on a similar product from the
other teacher. Similar academic achievement of students on
comparable assignments may result in differing grades, de-
pending upon the standards utilized by the teachers. No
research-based studies have been located to document the
reliability of grades in any form.

In addition to reliability, grades, as instruments of
measurement, should demonstrate validity if they are to as-
sess academic achievement appropriately (Erickson & Went-
ling, 1976; The National Commission on Excellence in Educa-
tion, 1983). The validity of grades is assessed by the ac-
curacy of predictions that can be based upon them. Stanley
and Hopkins (1972) have suggested that validity may be clas-
sified into three subtypes: 1) criterion related validity
(including predictive and concurrent validity), 2) content
validity, and 3) construct validity. To possess useful

validity according to this scheme, grades must accurately do the following: predict future performance, measure achievement, and match predictions of achievement obtained from other sources (Anderson, 1982). Unfortunately, insufficient data are available to suggest that grades can or cannot perform any of the aforementioned functions; their validity has not yet been established. No research studies have been located that have assessed the validity of grades.

The Psychometric Attributes of Marks

The psychometric attributes of marks may be analyzed according to the same scheme as was used to investigate the psychometric attributes of grades. The appropriate questions for this analysis become: "What is it that marks measure?" and "How well do marks measure whatever it is that they measure?"

In response to the first question (What is it that marks measure?), marks, for the same reasons as grades, should remain pure as measures of academic achievement. If marks are influenced by "non-academic" factors, they, like grades, lose their utility and become a nebulous compilation of elements; their meaning and consistency become compromised. An answer to the second question (How well do marks measure whatever it is that they measure?) focuses on an assessment of the relative reliability and validity of marks.

Three studies have been located which have assessed the internal consistency reliability of marks. Bendig (1953) assessed the internal consistency of information gathered to determine final marks in an introductory psychology course and found that the reliability of letter grades reflecting that information in various sections of the course ranged from .76 to .84. Fricke (1975) reviewed research findings on the internal consistency of mark averages earned in high school and college, and concluded that typical internal consistency coefficients were in the .80's.

Barritt (1966) sampled 250 cases from 1400 freshmen entering Indiana University in 1981, randomly split marks from each student's coursework in half, computed separate mark averages for each half, and then correlated the two halves. The average correlation coefficient was reported as .84.

Two authors have reported stability (test-retest types of reliability) estimates for marks. Odell (1950) reported estimates of the semester-to-semester reliability of college mark averages to be between .70 and .90. Terwillinger (1972) suggested that in high schools, the final marks students earn in courses are highly correlated with those obtained in previous courses of study, but specifics leading toward this conclusion were not reported.

According to the information reported on the reliability of marks, marks obtained by individual students may

demonstrate internal consistency, stability from year to year, and stability from course to course. Marks cannot be considered congruent, stable measures of achievement between groups of students, though, as evidence has not been found to support such comparisons.

Researchers have conducted validity assessments of marks. Their research has addressed efforts to establish predictive and concurrent validity estimates. Theoreticians, however, have only speculated upon the content validity of marks.

Studies assessing the predictive validity of marks suggest that marks possess a moderately high capacity to predict future educational performance (Davis, 1964; Hills, 1964; Lavin, 1965; Terwillinger, 1972). Davis (1964) reviewed studies of the predictive validity of marks and reported that when grade point averages for students in grades 10, 11, and 12 were compared to their grade point averages in college freshman classes, the typical correlation coefficients ranged from .55 to .65. Lavin (1965) also reviewed studies of high school grade point averages and concluded that grade point average is probably the best predictor of college success. Hills (1964) analyzed data from all of the publically-supported higher education institutions in Georgia for the five years from 1958-1962. Multiple correlations were used to assess the relationship between College Board scores, high school grade point

averages, and first year college grade point averages. The
average multiple correlation for the five years was .66.
Terwillinger (1972) sampled populations from two secondary
schools in Nashville, Tennessee, and correlated students'
marks with those obtained in previous years. Correlations
of .70 were reported, but a great deal of variability was
noted among different subject matter areas.

Attempts to establish the concurrent validity of marks
have suggested that marks correlate well with IQ as a pre-
dictor of classroom achievement, that they predict achieve-
ment as well as standardized tests, and that they correlate
acceptably with other marks obtained by a single student in
courses of the same subject matter. Terwillinger (1972)
correlated final average class marks to average class IQ in
his study of secondary schools in Nashville, Tennessee, and
reported a correlation of .60. Ahmann and Glock (1981)
summarized results from several studies of marks and con-
cluded that marks predict future educational performance as
well as scores on standardized tests, and that marks ob-
tained by individual students correlated acceptably to
marks the students earned in other courses of the same sub-
ject matter.

Research-based studies establishing the content valid-
ity of marks have not been located. Ahmann and Glock (1981)
have suggested that marks can demonstrate content validity
as measures of students' mastery of subject matter; however,

no research studies are reported to support this premise.

In summation, the available research indicates that marks for individual students can demonstrate predictive and concurrent validity characteristics. The content validity characteristics of marks have not yet been established by research findings.

THE LEGAL DOMAIN
OF GRADING AND MARKING SYSTEMS

Legal concerns in assigning grades and marks center around the principles inherent in two important legal issues: 1) the relationship of grades and marks to academic due process, and 2) the propriety of reducing grades or marks to effect sanctions for students' misconduct.

The Relationship of Grades and Marks
to Academic Due Process

Academic due process is an extension of the Due Process Clause of the Fourteenth Amendment of the U.S. Constitution to assure that no individual will be deprived of any life, liberty, or property interests without due process of law. Academic due process is embodied in the substantive and procedural protections that insulate students from unjustifiable deprivation of their rights by school authorities. The efficacy of the substantive and procedural safeguards sufficient to protect students' rights is defined by the courts, and this efficacy may vary depending

upon courts' interpretations of the issues at stake in cases of litigation. Thus, the substance and boundaries of academic due process protections are ultimately determined by the courts; the safeguards incorporated into these protections constitute the framework of academic due process.

Several court decisions address the relationship of grades and marks to academic due process. An outline of these decisions and their ramifications is presented in the following paragraphs.

A recent Supreme Court decision, <u>Board of Curators of the University of Missouri v. Horowitz</u> (1978), represents what may be considered a landmark decision relevant to the issue at hand. The Horowitz case, argued before the U.S. Supreme Court in 1977, examined the dismissal of a student, Charlotte Horowitz, from the University of Missouri Medical School as a result of her less than adequate academic performance. In this case, the Supreme Court suggested that schools could enforce academic standards. The ruling indicated that Ms. Horowitz could, as a result of academic performance deficiencies, be dismissed from medical school. Further, the Court's reasoning in the decision suggested that a different kind of due process concern exists for academic matters than that required for disciplinary matters as a result of <u>Goss v. Lopez</u>, (1975).

Jennings' (1979) review of the Horowitz decision suggested three practical provisions that incorporate the kind

of due process standard necessary for academic dismissals.
First, academic dismissals should be pursuant to published,
objective standards that specify a requisite mark average.
Second, a student should be provided a one term probation
or warning period before dismissal. Third, a student's
academic performance should be measured as unsatisfactory,
and a warning about the academic performance should be pro-
vided during the term in which the performance warning is
issued. These provisions set a pattern of reasonableness,
but they do not provide the course of formal notification
and hearing necessary to effectuate disciplinary dismissals.

Several lower court findings on academic dismissals
have suggested reasoning congruent with the Horowitz deci-
sion (Barnard v. Shelbourne, 1913; Foley v. Benedict, 1932;
Mustel v. Rose, 1968; Gaspar v. Bruton, 1975; Mahavongsanon
v. Hall, 1978), and these were cited by the Supreme Court
in rendering its opinion. Prior to the Supreme Court's de-
cision in Horowitz, several lower courts had already main-
tained that schools had the right to promulgate academic
standards. The Horowitz decision served to substantiate
reasoning offered by the lower courts.

Although the consummation of academic dismissals is
probably a moot point in public K-12 education, the ideas
highlighted in the previous paragraphs may be applied more
broadly to academic decisions in general. Schools possess
the authority to establish and maintain academic standards,

and they should not be reluctant to enforce these standards. When schools make decisions about students based on grades or marks, the decisions must be reasonable; however, a less restrictive course of notification and appeal than that required for disciplinary cases may be utilized.

A concurrent point in analyzing the relationship of grades and marks to academic due process is that courts have demonstrated a disinclination to intervene in educational decisions (National Association of Secondary School Principals, 1977). Several court cases support this contention (Connelly v. University of Vermont, 1965; Navato v. Sletten, 1976; Sandlin v. Johnson, 1981; Board of Education of the Hendrick Hudson Central School District v. Rowley, 1982). The consensus of these decisions has been that the courts have determined that educational decisions should be left to professional educators, because they, as professionals, possess the unique expertise necessary to make educational decisions. Courts have been reluctant to "second guess" decisions made by educators. Even though courts have been reluctant to review decisions made by professional educators, decisions that potentially interfere with the constitutional rights of students may be subjected to court scrutiny. Two cases illustrate this point.

The first case, Debra P. v. Turlington, 1981, examined the practice of using a competency test to determine eligibility for high school graduation. In this case, an appeals

court reasoned that a school decision could be considered open to judicial review if it legitimately interfered with a student's constitutional rights. The court reasoned that students held a property interest in receiving a high school diploma if they passed the courses required for high school graduation, and an additional requirement for receiving the diploma, such as passing a competency test consisting of content independent of that covered in high school coursework (the test did not demonstrate content validity) could constitute a violation of property rights.

The second case for consideration focused on a court's interpretation of a school district's decision to deny credit as a penalty for students' unexcused absences. In Anderson v. Board of Education, 1977, a Nebraska court reasoned that high school students possessed a property interest in receiving credits from classes if they demonstrated some reasonable measure of success. Again, the court was not reluctant to review educators' decisions, as these decisions impacted upon students' defined property rights.

Four of the court decisions reported above which have examined the relationship of grades and marks to academic due process have resulted from litigation in private schools and colleges. Direct application of the courts' reasoning to K-12 public education must be tempered. Decisions reported in this chapter suggest differing kinds of due process concerns for academic and disciplinary matters. These

decisions may set precedents potentially applicable to K-12 public schools, but until specific litigation evolves in the public school sector, conclusions based on these case findings must be applied with caution. The remaining court decisions on the relationship of grades and marks to academic due process do result from public school litigation, and a more straightforward application of their results is possible.

First, educational decisions are given high regard by the courts. Second, since educational decisions are given this high regard, the courts will probably continue to allow educators to make reasonable decisions about students and schools. Courts may choose to review these decisions, though, if they demonstrate the potential to interfere with students' constitutional rights. Decisions based on grades or marks may be subjected to court scrutiny if they can be linked to students' liberty or property interests. Third, if courts are willing to examine the content validity of tests which are used to make decisions interwoven with students' property rights, could they also analyze the content of grades or marks in the same fashion?

Finally, the U.S. Supreme Court has delivered a decision based on litigation from private industry which suggests that personnel assessment systems must demonstrate evidence of validity (State of Connecticut v. Teal, 1982). Can grades and marks, as assessments of students' academic

progress, be subjected to the same scrutiny that has been mandated for personnel assessments in industry?

Arguments may be proposed that schools can deny students access to specific educational tracks based on their performance. In accordance with these arguments, schools could dismiss students from certain courses of study based on their failure to maintain a pre-specified level of performance as measured by grades or marks. Decisions of this nature, with some built-in appeals, could probably satisfy procedural due process requirements. Nonetheless, courts may examine the decisions to dismiss or exclude students from programs, based on grades or marks, as violations of students' substantive rights, as these decisions could be interpreted as violations of students' liberty or property rights. Further, depending upon the courts' interpretations of the weight of interest at stake, more stringent due process protections may be required in some cases.

The Propriety of Reducing Grades and Marks to Effect Sanctions

There are a number of court cases which address the second legal issue for consideration, the propriety of reducing grades or marks to effect sanctions for students' misconduct. These cases will be outlined in the following paragraphs.

Three cases have been located which suggest that grade or mark reductions as sanctions for the misbehavior of

students are inappropriate. In 1965, the New Jersey Com-
missioner of Education, using his quasi-judicial powers,
ruled that the use of grades as deterrents or punishments
was inappropriate (Wermuth v. Bernstein, 1965). In ad-
dition, the Commissioner overruled a school board's policy
of assigning zeros for days of truancy or absenteeism due
to suspension, and then averaging these zeros with other
grades in the class (Minorics v. Board of Education of
Phillipsburg, 1972). In some instances, application of
this policy could cause a student to get a failing quarter-
ly course mark for even a single absence. A major aspect
of this action seems to have been the severity of punish-
ment rather than the objection to the grade penalty itself
(National Association of Secondary School Principals, 1977).
Last, in 1975, a Kentucky appeals court examined a regula-
tion by which a student's grade was reduced as a result of
unexcused absences. The court determined that this grade
reduction represented an additional punishment, beyond the
unexcused absence, which the school should not levy (Dorsey
v. Bale, 1975).

Two Illinois courts have suggested a somewhat differ-
ent line of reasoning on the issue of grade reductions. In
1976, an appeals court upheld a school district's policy of
lowering a student's grade one letter grade per class for
each unexcused absence, concluding that this was a reason-
able and rational regulation to combat truancy. This

sanction was not determined so harsh as to deprive students of substantive rights (Knight v. Board of Education, 1976). In 1978, a different appeals court examined the reasoning in Knight in rendering a decision on a similar matter. This appeals court agreed that grade reductions did not represent denial of students' substantive rights (Hamer v. Board of Education, 1978). A reasonable conclusion seems to be that when courts do not view grade reductions as harsh or as potential violations of students' substantive rights, they may potentially be held as valid.

A county court decision rendered in Missouri provides one last directive for examination. In this case, a student was given failing grades in concert choir and band for receiving an unexcused absence from a required concert. The student charged that this grade reduction was unfair, even though prior notice was provided which indicated that any unexcused absence from a performance would result in an "F". The reasoning of the county court judge suggested not only that the student was made aware of and given reasonable notice of the course requirements before his unexcused absence, but that participation in specified activities could be held as a valid course requirement; failure to complete this participation could trigger a grade reduction (Johnson v. Shineman, 1982).

Court tests of the efficacy of reducing grades as an academic sanction center around punishments for student

absences, truancy, or misbehavior in school. The court
tests have produced mixed results. Several cases support
the contention that it is illegal to reduce students'
grades as sanctions: <u>Wermuth v. Bernstein</u>, 1965; <u>Minorics
v. Board of Education of Phillipsburg</u>, 1972; and <u>Dorsey v.
Bale</u>, 1975 (see also Dressel, 1976; Valente, 1980). Other
cases indicate that grade reductions may be legitimately
used as academic sanctions under some circumstances (<u>Knight
v. Board of Education</u>, 1978; <u>Johnson v. Shineman</u>, 1982).

Implications of the litigation relevant to reducing
grades as sanctions must be appraised in light of two con-
texts: the level of court in which the decision is ren-
dered, and the state in which the litigation was generated.
The cases which have been decided to date have all been
tried at a level no higher than district or state appellate
court; no cases have yet gone to the Supreme Court level.
Thus, the scope and applicability of the findings in these
cases is somewhat limited and conflicting decisions have
been promulgated. Further, the court decisions on this is-
sue have been confined to action in several states, and the
findings in these cases may be limited specifically to the
few states in which the litigation has transpired.

According to Nolte's (1980) review of litigation, the
practice of reducing grades to effect sanctions can unfair-
ly prejudice a student's substantive due process rights.
This prejudice results because a diminished grade is

assigned which does not accurately reflect the student's academic performance. For example, a student is given a zero as a class daily grade for truancy, then is given a make-up test. Since the zero is weighted against the results of that test, an inappropriate grading practice results. The zero dilutes the measurement of the student's academic performance, as it is averaged with the test grade, and a diminished grade is assigned that does not represent the student's actual performance. This reasoning is congruent with that proposed earlier in this chapter which suggests that grades and marks should reflect academic achievement only. In addition to creating legal concerns, grade reductions of the type illustrated here represent poor psychometric practices.

The National Association of Secondary School Principals published a legal memorandum in October of 1977 summarizing litigation on grade reduction. The four conclusions stated in this memorandum remain consistent with recent litigation and provide guidelines for educational practice. These four conclusions are offered as a final summary statement. First, grade reductions as means of discipline for student actions unrelated to academic performance may be open to challenge either on constitutional grounds or for being beyond the bounds of reasonableness. Second, automatic grade (or mark) reductions for truancy will usually be permitted if they can be shown to have a

reasonable educational relationship to the offense. Third, the issuance of low grades, refusal of credit, or dismissal for academic as opposed to disciplinary reasons, are actions unlikely to be subject to review by the courts unless there appears to be a clear basis for alleging that the action by school authorities was arbitrary, capricious, or in some other way an abuse of discretion. Fourth, in general, the more severe the effect of the academic penalty on a student, the more likely that due process requirements may apply.

THE LEARNING DOMAIN
OF GRADING AND MARKING SYSTEMS

The concepts of reward and punishment are deeply rooted in the behavioristic traditions of psychology, and suppositions inherent in these concepts may be applied to grades and marks to appraise their value as rewards and punishments. There is an extensive body of literature delineating the postulates of reward and punishment theories, but a thorough review of this literature is beyond the scope of this chapter. Therefore, the ensuing discussion is intended to be illustrative rather than exhaustive.

Classic research studies conducted by Thorndike (1911), Pavlov (1927), and Skinner (1938) have evaluated the effectiveness of rewards in changing behavior. Conclusions from these studies have been refined by contemporary researchers to suggest the following definition for reward: a reward

is any occurrence which follows a behavior that tends to strengthen the probability that the behavior will reoccur (Mussen, Conger, & Kagan, 1963; Hilgard & Bower, 1966). Rewards are events or tangible objects that are provided as consequences of a specified behavior with the intent that they will encourage that same behavior to reoccur.

Rewards may be classified as either primary or secondary (Hilgard & Bower, 1966). Primary rewards are inherently worthy and correspond with primary physiological needs. Examples of primary rewards are food, water, sleep, and warmth. Secondary rewards (sometimes called learned rewards) do not possess inherent worth. They acquire their reward value because they have been associated with the gratification of some motive. Examples of secondary rewards are praise, social approval, and money.

Grades and marks may be classified as secondary rewards. For them to earn value as rewards, as is required by definition, they must become associated with the gratification of some motive. For example, a student is motivated to complete tasks assigned in a course of study with a high level of accuracy. When the assigned tasks are completed with the expected level of accuracy, good grades result. Eventually, high marks are earned for acceptable levels of academic performance. Theoretically, the grades and marks can become meaningful to the student because they have been associated with the motive to complete tasks with

a high level of accuracy.

Once grades and marks have obtained reward value, their issuance should strengthen the student's behavior. Students who receive high grades and marks should, by theory, behave in such a way that the actions used to gain the rewards will be repeated. Students should continue to exhibit the relevant behaviors that resulted in the high grades and marks. If a student has obtained good grades and marks, probability suggests that the student will work to earn the grades and marks in future endeavors.

There is evidence available to suggest that grades and marks have acquired value as rewards for some students. Marks obtained by an individual student have been found to be highly correlated with those obtained by the student in previous courses of study (see the discussion on the psychometric properties of marks). Although research has not been located to evaluate a similar paradigm for grades, the individual grades which make up composite marks probably also correlate well with those obtained previously by individual students. Grades, like marks, have probably acquired reward value for some students.

Grades and marks are not likely to have acquired reward value for all students, though. There is evidence to suggest that grades and marks do not lead toward increased motivation of all students. Cullen et. al. (1975) reported differential effects for grades used as positive incentives and grades used as negative incentives. Further, in the

Yarborough and Johnson (1980) study cited in the section
on problems associated with grades and marks, the authors
concluded that marks meant different things to different
students. Broadfoot (1979) has also postulated that grad-
ing alienates students and does not provide for increased
student motivation.

Mitzel, Best, and Rabinowitz (1982) have explained the
differential motivation effects of grades and marks through
exploring the constructs of intrinsic and extrinsic motiva-
tion. Intrinsic motivation relates to behavior that is
self-perpetuating or performed without external rewards.
For example, students who study predominantly because they
want to learn course material are considered to be intrin-
sically-motivated. Extrinsic motivation relates to behav-
iors performed because a student hopes to receive certain
rewards. Extrinsically-motivated students study primarily
to earn high grades and marks or to avoid low ones, where-
as intrinsically-motivated students study for other, inter-
nally-based reasons. Extrinsically-motivated students may
be motivated by grades and marks, while it is unlikely that
intrinsically-motivated students will be motivated by high
or low grades and marks.

The available information suggests that grades and
marks may demonstrate the potential to serve as rewards for
some students. Evidence does not suggest that they have
acquired reward value for all students.

Punishment, like reward, has been an object of psychological research for many years. Punishment is defined as any consequence following a behavior which reduces the probability of that behavior recurring (Walters & Grusec, 1977). Examples of punishments are: spankings, groundings, "time outs", and monetary fines. Experts are not in complete agreement about the efficacy of punishment. Walters and Grusec (1977) have suggested that punishment, by itself, does not produce permanent effects, and Cronbach (1963) has proposed that mere non-reinforcement eliminates behaviors more permanently than punishment.

Travers (1977) reviewed the literature relevant to punishment and concluded that it could be considered effective under some circumstances. The weight of the research reviewed by Travers (1977) indicated that very mild punishment did little to inhibit behavior. As punishment became stronger in intensity, it became capable of temporarily suppressing responses. Still stronger punishment produced a more marked suppression with some lasting effects. Very strong punishment produced a very complete and very permanent suppression of the punished response.

Arguments may be proposed that punishment should or should not be used in schools, and there is probably no empirical resolution to satisfy either argument. Decisions on whether or not punishment should be used probably result from philosophical, not empirical, choices.

Researchers agree that when punishment is used, it should be related to the nature of the misbehavior (Klausmeier & Goodwin, 1966; Hurlock, 1978). Gardner (1983) has suggested that educators have been granted broad legal authority to administer punishments to students, and that courts are probably reluctant to intervene as long as the punishments are reasonable and supported by sound justifications. Logically then, the nature and severity of a punishment should be related not only to the behavior, but to the degree of transgression promoted by the behavior.

Grades and marks can represent punishments in two ways. First, the issuance of low grades or marks can connote punishment for lack of achievement, as they are issued because students have not maintained the expected standard of academic achievement. Second, the reduction of grades or marks to effect sanctions for students' misconduct (as outlined in the section on legal concerns in assigning grades and marks) can also represent a form of punishment (Anderson, 1982).

There is a dearth of research related to the use of grades and marks as punishers. Information reported by Cullen et. al. (1975) and Mitzel, Best, and Rabinowitz (1982) illustrates the value of grades and marks as punishers. Cullen et. al. (1975) suggested that the threat of a low mark may elicit increased task completion by some students, but that all students may not be motivated by such a

threat. Mitzel, Best, and Rabinowitz (1982) suggested that extrinsically-oriented students may be motivated by the use of low grades and marks as punishers, but low grades and marks may not motivate intrinsically-oriented students. The information reported here, although brief, suggests that grades and marks may not be perceived as punishers for all students.

SUMMARY

Educators agree that grading and marking systems are necessary to monitor, chart, and report students' academic growth systematically. Schools have been given wide latitude to develop and implement a variety of grading and marking systems, and the grades and marks generated by these systems have become prominent standards in assessing student achievement.

Researchers do not agree upon the merits or liabilities of one particular type of grading or marking system, and many different types of systems with varying compositions are presently utilized in our schools. Research studies suggest that the traditional systems which have been employed to derive grades and marks are counterproductive and ineffective, yet they continue to persist as mediums to report students' progress.

The solution to this problem is not to develop a totally new system, but to revise and validate the ones which

already exist (Ebel, 1979). School administrators should design, revise, and validate grading and marking plans responsive to their district's needs. The types of grading and marking systems which have been outlined in this chapter provide models for the general types which can be considered for development.

Grades and marks derived by any system should reflect academic ahcievement only, and evaluative assessments of students' attitudes, efforts, personalities, attendance, and behavior should not be considered as factors in their determination. Evaluative assessments of non-academic factors may be reported concurrently with grades and marks by the use of checklists or descriptive word statements.

There is insufficient information to support the contention that grades are sufficiently reliable or valid. The available information indicates that grades are quite unreliable across students. This unreliability may be attributed to two factors. First, teachers, schools, and school districts utilize differing standards to assign grades. Similar achievement on assignments in the same courses of study as taught by different teachers may result in different grades. Second, teachers do not all derive grades similarly. Some teachers may grade according to fixed academic standards, others may base grades on variable standards, while still others may base grades on a combination of academic and evaluative factors. Grades lose some degree of

comparability due to this variety of inputs.

Marks, however, have demonstrated reliability as measures of the performance stability of individual students from year to year and from course to course; they probably possess sufficient reliability to serve as useful estimates of the academic performance stability of individual students. Marks, like grades, have not demonstrated comparability between students, and reliable comparisons between students' marks are not possible.

Marks have demonstrated a moderate degree of predictive validity; they can serve as useful predictors of the future academic performance of individual students. Marks have also demonstrated a moderate degree of concurrent validity as they predict individual achievement as well as IQ, standardized tests, and other marks earned by students. Nonetheless, the content validity of marks remains suspect in most settings, as research-based information has not been located to establish such validity. Further, several authors have offered detailed methods for appropriately deriving marks (Davis, 1964; Brown, 1971; McIntosh, 1974; Chase, 1978; Ebel, 1979; Gronlund, 1981). The methods offered by these authors involve completing score transformations to create standard scores, and an understanding of terms such as variance, standard deviation, mean, and mode. The ability of many teachers to understand the relevant terminology and apply the proper score transformations is

suspect. Improper applications of methods for deriving marks may inhibit the content validity of marks to an unknown degree.

Implications of the legal and learning concerns encountered in assigning grades and marks are as follows. School districts must develop and apply a reasonable, written process to make decisions based on students' grades or marks. These decisions should be pursuant to published, objective standards of which students and parents are appraised at the outset of each course of study. Parents of students, as well as the students who encounter academic problems should be notified, in writing, of academic deficiencies during the term they become evident.

Schools must assure that grades and marks demonstrate reliability and validity, as grades and marks may be used to facilitate decisions about students which convey serious and far-reaching consequences. With respect to grade reductions as sanctions for student conduct, the penalties created by such grade reductions may exert an improper influence upon the final marks that a student obtains, and the distorted final mark a student earns may impact negatively upon future college, vocational, or career opportunities. Courts' analyses of grade reductions as sanctions for student conduct suggest that, at least potentially, only severe grade reductions will be considered as violations of students' rights. Minor types of reductions,

especially if they are rationally related to student conduct consistent with appropriate academic achievement, or if they are specified as related to class participation as a part of academic performance, may be held as legitimate. Major grade or mark reductions as sanctions, though, will probably be indicted as inappropriate.

Courts may approve grade or mark reductions as sanctions for behavior which bears a clear relationship to relevant educational standards. Furthermore, grades and marks can be legitimately reduced for students' failure to achieve the academic standards promulgated by schools.

Despite the potential legality of grade and mark reductions, their effectiveness is questioned on theoretical principles for the following reasons. First, for grades and marks to serve as punishments, they must acquire worth as meaningful symbols to students. Since grades and marks have not been proven useful as tools for increasing the motivation of all students, they are probably not meaningful to all students. Withholding a high grade or awarding a low grade or mark as a punishment for failure to achieve certain academic standards will probably not impact consistently on all students.

Second, grade or mark reductions as sanctions for student misbehavior violate the standard that punishment should be related to the transgression. A grade or mark reduction represents an academic penalty for a behavioral

transgression. Third, the minor types of grade (or mark) reductions which have been approved by the courts probably carry little weight as punishments. These are likely to have minimal consequences, and punishments with such slight consequences have been found to be ineffective in changing behavior.

Last, the issuance of low grades or marks can connote punishment for lack of achievement when students have not mastered the expected level of academic performance. There is not sufficient evidence to conclude that the issuance of low grades or marks creates any significant long term learning effects. Low grades or marks should be issued to represent levels of student achievement commensurate to their value, but the issuance of low grades or marks should not be used as a threat to students, as inconsistent student effects may result.

In summation, grades and marks should remain prominent as standards of students' academic achievement. Schools should not hesitate to use grading and marking systems to designate levels of achievement and to enforce academic standards, but care must be taken to assure that the systems used to derive grades and marks are reliable and valid.

CHAPTER III

DESIGN OF THE STUDY

INTRODUCTION

The purpose of this study was to create a pool of test items to assess knowledge of principles and practices associated with the rational, psychometric, legal, and learning domains of grading and marking systems, for which reliability estimates and content validity characteristics would be established. In this chapter, research procedures are described. The chapter is divided into the following sections: 1) Research Objectives, 2) Population, 3) Sample, 4) Development of the Test Items, 5) Data Gathering Procedures, and 6) Data Analysis Procedures.

RESEARCH OBJECTIVES

Two research objectives were generated to assist in validating the test items. These were:

1. To ascertain the extent of agreement among experts, with content knowledge congruent with the four domains of grading and marking systems, as to what information is essential for practitioners to either know or apply to implement grading and marking systems that are consistent with recommended practices.

2. To discover the extent of agreement among teachers who regularly apply grading and marking systems as to what information is essential

to either know or apply to imple-
ment recommended grading and mark-
ing systems.

POPULATION

Two separate populations, one for each of two test item validation phases, were used to complete the study. The population for the first phase was the aggregate of professionals from four fields of study: 1) school psy-chology, 2) educational psychology, 3) measurement and evaluation, and 4) school law. The population for the sec-ond phase consisted of the teachers listed in the 1983-1984 edition of the Nebraska Educational Directory.

SAMPLE

Samples were chosen from each of the populations to comprise the respondent groups for the two phases of test item validation. The sample utilized in the first valida-tion phase was derived by implementing a peer nomination plan.

A random numbers table was used to select 20 names from each of three membership divisions of the American Psychological Association, as listed in the 1983 edition of the American Psychological Association Directory: Division 5 (Evaluation and Measurement), Division 15 (Educational Psychology), and Division 16 (School Psychology). The ran-dom numbers table was also used to select 20 names from the

1984 roster of the Board of Directors of the National
School Boards Association Council of School Attorneys.
Each of the 80 professionals identified by this process
was asked to nominate at least five nationally known ex-
perts from their respective fields to participate in the
validation study.

From the lists of nominees provided by the original
80 professionals, 10 from each professional group were se-
lected for inclusion in the validation study. In order to
enlist 10 experts from each group of nominees, the most
frequently nominated individuals were chosen first. Then,
the next most frequently nominated persons were chosen, and
so on, until each of the nomination lists was left with in-
dividuals who had only a single nomination. The remaining
nominees within each group were assigned numbers, and a
random numbers table was used to draft individuals until
10 persons were nominated from each group. These 40 ex-
perts comprised four groups of 10 professionals, each rep-
resenting an area of expertise congruent with one of the
four domains of grading and marking systems.

The sample for the second test item validation phase
consisted of 250 Nebraska teachers. Teachers who comprised
the sample were randomly selected from those listed in the
1983-1984 edition of the Nebraska Education Directory by
following a two-step process. First, a series of page num-
bers was chosen by drawing numbers from a random numbers

table. Second, every fifth name from the second column of name listings on each of the designated pages was selected for participation in the study until the desired sample size was attained.

Names of administrators and supervisors were disqualified for the purposes of the study. When the name of an administrator or supervisor was picked, it was automatically dropped from the sample. The final composition of the sample for the second validation phase was as follows:

Elementary School Teachers - 124 Teachers
Middle School Teachers   - 13 Teachers
Jr.-Sr. High School Teachers - 113 Teachers

Areas of endorsement represented in the sample, and the number of teachers with each endorsement, were as follows:

Art -	10	Health, P.E. -	18
Music -	20	Special Education -	16
English Language -	20	Media & Library -	6
Science -	8	Business -	13
Counselor -	5	Home Economics -	6
Social Service -	9	Chapter I	7
Agriculture -	2	Industrial Arts -	9
Math -	13	Foreign Language -	6
Elementary Education -	82	None Listed -	2

DEVELOPMENT OF THE TEST ITEMS

The test items were designed and validated by implementing a plan created to establish reliability and content validity. Material for inclusion in the items was determined from a content analysis of research findings reported in Chapter II. Information to be delimited was sampled

from research findings associated with the rational, psychometric, legal, and learning domains of grading and marking systems, but no attempt was made to sample each of the domains equally, as differing amounts of information were found to exist across domains.

Once the material to be included in the test had been outlined, a format for the items was selected. A multiple-choice format was chosen, as multiple-choice items have exhibited a high degree of utility. Multiple-choice items have demonstrated the potential to sample a wide variety of data sources, they are easily administered and scored, and they yield information that can readily be subjected to statistical scrutiny (Nunnally, 1967).

Item stems, answers, and foils were constructed to represent the information which had been sampled from the research findings reported in Chapter II. Correct answers were randomly assigned a position among the foils to control for response position variables. This randomization was completed by first assigning a correct response to each possible response position through the first four items in chronological order, and then repeating this sequence as often as necessary to assign correct response positions to all of the stems. Items in which "all of the above" and "none of the above" responses represented the correct choices were exempted from this process (some items were given "all of the above" or "none of the above" choices as

foils). Two additional items were also exempted from the
randomization, as they both were given five answer choices.
Correct answers for these items were assigned to response
position five. After the randomization of correct answers
and foils was completed for all of the items, each response
was given a number from one to five.

The finalized test items were assigned to a sequence
as follows. Each item stem was given a number from 1 to
22, depending on the order in which they were created.
These numbers were used to determine a final sequence by
ordering the test items in the sequence in which their
original numbers occurred in a random numbers table. After
all of the test items were assigned to this final sequence,
item numbers were reassigned to reflect the final order.

Once the completed test items were assigned to their
final sequence, they were reviewed by three University of
Nebraska-Lincoln staff members. Each of these persons pos-
sessed an earned doctorate in education, and all had con-
structed and employed tests in their work. These individ-
uals provided feedback on the clarity of the stems, an-
swers, and foils, the length of the test instrument, and
its general organization. The clarification of several of
the foils and some minor grammatical changes were made as
a result of this feedback.

A two-phase validation study was completed as the fi-
nal step in developing the test items. The purposes of

this validation study were to enhance content validity and
to obtain data for reliability assessments. Two separate
populations were sampled, one for each phase of the study.

Phase I

The first phase of validation was designed to enhance
the content validity of the test items by gathering data
from a professional jury regarding item content, nature,
and merit. A jury comprised of experts with knowledge con-
gruent with the four domains of grading and marking systems
was selected to assist with this phase of validation.

Each juror was provided with a cover letter, sample
test with the correct answers designated for each item, a
validation worksheet, and a self-addressed, stamped return
envelope (see Appendix A). The jurors were asked to com-
plete three tasks: 1) sort the items into two categories,
items that measure knowledge of the attributes of grades
and marks, and items that measure knowledge of recommended
principles for issuing grades and marks to students,
2) list the items that represent information essential for
educators to either understand or apply to insure good
grading and marking practices, and 3) list any concepts or
issues not included in the test that should be included.

Phase II

The second phase of validation was designed not only
to collect further information on the merit of the test

items, but also to provide data for reliability estimates and item analyses. A sample of Nebraska teachers served as data sources for this second validation phase.

Each teacher was provided a cover letter, a sample of the test items, a self-addressed, stamped return envelope, and validation worksheet (see Appendix A). Respondents were asked to complete three tasks: 1) designate the correct answer for each test item, 2) list the items representing information essential for practitioners to either know or apply, and 3) list any concepts or issues not included in the test that should be included.

After the test development plan was completed, data analyses were conducted to appraise the worth of the items. A final pool of the test items with the best psychometric properties was derived by appraising the results of data analyses, but further validation study will be needed before a complete pool of items is realized. Means for these analyses are reported later in this chapter; results of the analyses are reported in Chapter IV.

DATA GATHERING PROCEDURES

Data were collected by using qualitative and quantitative methods during the various stages of the test development plan. Several kinds of information were obtained. Some of the information which was collected consisted of open-ended, language-mediated responses which had to be

analyzed qualitatively, e.g. "Are there any concerns or
issues not included in the test items that, in your opin-
ion, should be included? If so, please list them." In-
formation was also obtained from closed-choice responses
which were subjected to quantitative analyses, e.g. "List
the numbers of test items that represent pieces of informa-
tion you deem essential for educators to either understand
or apply to insure good grading and marking practices."

Verbal and written comments obtained from critiques of
the item stems, answers, foils, the length of the test, and
its general organization were provided not only from Uni-
versity of Nebraska-Lincoln staff members by design, but
also from respondents to the tasks assigned during the two
validation phases. Each of the three university staff mem-
bers was specifically asked to provide either written or
verbal feedback, and this was completed during one of the
early steps of item development work. Further, some of the
validation respondents wrote comments regarding specific
organizational, grammatical, or informational concerns.
This feedback was obtained during the latter stages of test
development.

Each of the persons chosen at random to assist in
identifying a professional jury for the first validation
phase was sent a cover letter, worksheet, and a self-ad-
dressed, stamped return envelope. These persons were asked to
list the names of professionals they perceived to be experts

in their respective fields (see Appendix A). After two weeks, non-respondents were sent a duplicate information packet. Data was collected from the professional jury and the Nebraska teachers by the methods outlined in the section on development of the test items (see Appendix A).

DATA ANALYSIS PROCEDURES

Qualitative analyses of the information collected on stem-foil matches, grammar, punctuation, and the overall organization of the test items were conducted before the finalized set of test items was sent to the professional jury members. Each verbal or written response related to this information was weighed, and necessary changes were made when the information was perceived to warrant such changes. In addition, qualitative responses from jury members and teachers regarding the kinds of information contained in the test items, as well as unsolicited remarks about the worth of the items, were collected during the two phases of item validation. These provided insight into why teachers performed as they did on the test items, and can be weighed during future modifications of the test items.

Quantitative analyses were conducted as follows. Percentage transformations were used to describe information about the frequency with which test items were categorized as representing knowledge or practice. Percentages were also used to describe the frequency with which test items

were designated as representative of information essential to know or apply to implement recommended grading and marking practices. Number and percent scores representing information about items answered correctly were used to illustrate teachers' performance on the test items and to evaluate the worth of the items. Percentages of teachers' responses to each answer choice were used to evaluate the worth of answers and foils.

Computerized data analyses to assist with the item analyses and appraisal of the items were accomplished at the Nebraska Evaluation and Research (NEAR) Center on the University of Nebraska-Lincoln campus. Pearson product-moment correlation coefficients were used to assess relationships between jury member designations of essential test items, teachers' designations of essential items, and teachers' correct responses to the items. The acceptable level of confidence for determining whether or not these correlations were unequal to 0 was .05. Means, standard deviations, estimates of kurtosis, and estimates of skewness were calculated to describe the distribution of jury members' and teachers' designations of essential items and teachers' correct responses to the test items.

A t-test for dependent groups was conducted to assess differences in teachers' performance on essential items and non-essential items. The acceptable error estimate was set at the .05 level. Means of teachers' scores on essential

knowledge and essential practice items were compared to assess differences in performance on these two types of items.

The Phi coefficient was used to compute score correlations between items. The point-biserial formula was used to calculate correlations between scores on individual items designated as representations of essential information and total scores based on aggregate performance on all essential items.

CHAPTER IV

RESULTS OF DATA ANALYSES

INTRODUCTION

The purpose of this study was to create a pool of test
items to assess knowledge of principles and practices asso-
ciated with the rational, psychometric, legal, and learning
attributes of grading and marking systems, for which relia-
bility estimates and content validity characteristics would
be established. The test items were developed by implement-
ing the test item development plan outlined in Chapter III.

Information to complete two research objectives was
collected to assist in validating the test items throughout
the implementation of the test development plan. These were:

> 1. To ascertain the extent of agree-
>   ment among experts, with content
>   knowledge congruent with the four
>   domains of grading and marking
>   systems, as to what information
>   is essential for practitioners to
>   either know or apply to implement
>   grading and marking systems that
>   are consistent with recommended
>   practices.
>
> 2. To discover the extent of agree-
>   ment among teachers who regularly
>   apply grading and marking systems
>   as to what information is essen-
>   tial to either know or apply to
>   implement recommended grading and
>   marking systems.

Samples from two populations were used during the fi-
nal stages of the test development plan to complete tasks

designed to assess reliability and enhance content validity. A professional jury consisting of 40 persons possessing knowledge congruent with attributes of the four domains of grading and marking systems was asked to complete tasks assigned during the first of these stages; a group of 250 Nebraska teachers was asked to complete tasks during the second stage. The professional jury was derived by using a peer nomination plan. The sample of Nebraska teachers was obtained by implementing a plan in which teachers were randomly selected for the study. Response rates obtained from the samples, and response rates secured from persons nominating professionals for the jury, are listed in Table 1.

TABLE 1

Sample Response Rates

Sample	Responses Received/ Responses Possible	Rate of Response
Professional Jury Nominator Totals	40/80	50%
Totals by Subgroup:		
School Psychologists	16/20	65%
Educational Psychologists	8/20	40%
Evaluation and Measurement Psychologists	9/20	45%
School Attorneys	10/20	50%
Professional Jury Totals	19/40	48%
Totals by Subgroup:		
School Psychologists	4/10	40%
Educational Psychologists	6/10	60%
Evaluation and Measurement Psychologists	7/10	70%
School Attorneys	2/10	20%
Nebraska Teachers	167/250	67%

QUALITATIVE DATA ANALYSES

Qualitative comments regarding the composition and content of the test items were weighed on an individual basis and were used to refine the test items into their final validation form (see Appendix B).

QUANTITATIVE DATA ANALYSES

Professional jury respondents were asked to specify test items which represented information they perceived to be essential for practitioners to either know or apply to implement recommended grading and marking systems. The frequencies with which items were designated as representing such information are reported in Table 2. Percentages of designation listed in Table 2 were delimited by comparing the number of times an item was specified as representing essential information to the total number of times an item could be designated as representing essential information, as determined by the total number of jurors responding to this task.

TABLE 2

Items Designated Essential
by the Professional Jury

Item Number	Frequency of Designation
1	.25
2	.17
3	.17
4	.50
5	.33
6	.17
7	.17
8	.50
9	.67
10	.33
11	.33
12	.67
13	.08
14	.67
15	.50
16	.50
17	.58
18	.58
19	.67
20	.67
21	.58
22	.50

The Nebraska teachers were also asked to specify test items perceived to represent information essential to know or apply to implement recommended grading and marking systems. Designation frequencies were delimited by comparing the number of times an item was specified as representing essential information to the total number of times an item could be specified as representing essential information, as determined by the total number of teachers responding to this task. Frequencies with which test items were

designated as representing essential information by the
teachers are listed in Table 3.

TABLE 3

Items Designated Essential
by Nebraska Teachers

Item Number	Frequency of Designation
1	.36
2	.40
3	.35
4	.42
5	.24
6	.40
7	.39
8	.40
9	.60
10	.33
11	.33
12	.25
13	.32
14	.44
15	.50
16	.32
17	.39
18	.33
19	.36
20	.65
21	.60
22	.31

Raw scores representing the total number of items each
of the Nebraska teachers answered correctly are reported in
Table 4. Scores ranged from 1 to 14 correct on the 22 item
test, and the modal correct score was 10.

TABLE 4

Raw Test Scores Obtained
by Nebraska Teachers

Number of Test Items Answered Correctly	Number of Teachers Obtaining Each Score
14	1
13	6
12	3
11	16
10	30
9	22
8	28
7	29
6	14
5	8
4	7
3	2
2	0
1	1

An item by item breakdown of performances which lists
the percentage of teachers passing and failing each item
is outlined in Table 5. Performances ranged from only 2%
correct on item 12 to 80% correct on item 21. The mean
correct score across all 22 items was 38%.

TABLE 5

Item by Item Performance of Teachers

Item Number	Percent of Teachers Passing	Percent of Teachers Failing
1	17	83
2	3	97
3	19	81
4	42	58
5	72	28
6	4	96
7	4	96
8	16	84
9	56	44
10	26	74
11	33	67
12	2	98
13	33	67
14	69	31
15	45	55
16	29	71
17	35	65
18	40	60
19	63	37
20	72	28
21	80	20
22	71	29

Table 6 presents a percentage summary of teachers' choices on each of the response positions for the 22 items. Several items had only four answer choices; therefore, an x has been placed in response choice number 5 for those items.

TABLE 6

Teacher Response Choices in Percentages

Item Number	Correct Choice Number	Response Choices				
		1	2	3	4	5
1	2	14	17	9	58	2
2	1	3	1	36	40	20
3	4	46	30	5	19	x
4	1	42	23	17	14	4
5	4	6	16	6	72	x
6	2	78	4	4	14	x
7	1	4	31	31	34	x
8	5	22	30	26	6	16
9	5	6	3	32	3	56
10	3	1	2	26	63	8
11	4	42	21	4	33	x
12	4	11	36	16	2	35
13	1	33	21	8	36	2
14	3	0	9	69	8	14
15	1	45	0	0	2	53
16	4	30	30	11	29	x
17	2	41	35	8	13	3
18	4	11	30	19	40	x
19	2	31.	63	5	1	x
20	5	4	15	3	6	72
21	5	6	0	0	14	80
22	5	15	13	0	1	71

Table 7 depicts a listing of the test items, the percent of jury members and teachers who designated each item as representing information essential for practitioners to know or apply, and the percent of teachers who passed each item.

Pearson product-moment correlation coefficients were computed between each of the columns of percentages reported in Table 7. Results of those correlations are listed in Table 8. Probabilities that the coefficients represented some degree of relationship between the data sources are

also listed. Only the coefficient which assessed the relationship between jury designations of essential items and correct teacher responses was significant at or beyond the .05 level of confidence, which suggested a 95% chance that there was some relationship between these two sets of data.

TABLE 7

Jury and Teacher Designations of Essential Items
and Percentages of Teachers Passing Each Item

| Item Number | Essential Designations | | Correct Responses |
	Percent of Jury	Percent of Teachers	Percent of Teachers
1	25	36	17
2	17	40	3
3	17	35	19
4	50	42	42
5	33	24	72
6	17	40	4
7	17	39	4
8	50	40	16
9	67	60	56
10	33	33	26
11	33	33	33
12	67	25	2
13	8	32	33
14	67	44	69
15	50	50	45
16	50	32	29
17	58	39	35
18	58	33	40
19	67	36	63
20	67	65	72
21	58	60	80
22	50	31	71

TABLE 8

Correlation Coefficients Between Jury Designations
of Essential Items, Teacher Designations
of Essential Items, and Teacher Correct Responses

	Jury Designations	Teacher Designations	Teacher Correct Responses
Jury Designations	1.00		
Teacher Designations	.37 p=.088	1.00	
Teacher Correct Responses	.56 p=.006	.40 p=.067	1.00

Table 9 illustrates distributions for jury designations of essential items, teacher designations of essential items, and teacher correct responses when they are plotted according to percentage ranges. Analyses of the percentage distributions of jury designations of essential items, teacher designations of essential items, and teacher correct responses indicate the following. The mean percentage with which test items were designated as representations of essential knowledge by the professional jury members was 43.6. The standard deviation of the distribution of jury members' designations was 19.9. Kurtosis of the distribution was -1.3, and skewness was -.4. The measures indicate that the distribution was bimodal, flat, and relatively symmetric.

TABLE 9

Distribution of Jury Designations of Essential Items,
Teacher Designations of Essential Items,
and Correct Responses to Items by Teachers

| Percentages of Designation or Correct Responses | Number of Designation Percentages | | Correct Responses |
	Jury	Teachers	Teachers' Responses
75-80			1
70-74			3
65-69	5	1	1
60-64		2	1
55-59	3		1
50-54	5	1	
45-49			
40-44		5	2
35-39		5	2
30-34	3	6	2
25-29	1	1	2
20-24		1	
15-19	4		3
10-14			
5-9	1		
0-4			4

The mean percentage with which test items were desig-
nated as representations of essential information by the
Nebraska teachers was 39.5. The standard deviation of the
distribution was 10.8. Kurtosis was -.8, and skewness was
1.1, which suggested the distribution was unimodal, flat,
but not as flat as for the jury, and positively skewed.

The mean percentage with which items were answered
correctly by the Nebraska teachers was 37.8. The standard
deviation of the distribution of teachers' scores was 25.4.
Kurtosis was -1.2, and skewness was .2, which suggested

that the distribution was flat, elongated, and relatively symmetric.

Inspection of the distribution of jury members' designations of essential items suggested a natural breaking point at the 50% designation rate. Two separate distributions, one above the 50% designation rate and one below the 50% designation rate were apparent, thus suggesting two different types of responses to the items. Items at or above the 50% designation rate were considered to be the jury's representations of essential information, whereas items below the 50% designation rate were considered representations of non-essential information.

Analysis of the distribution of teachers' designations of essential items indicated that there was no clear distinction made between the designations of essential and non-essential items. The distribution was positively skewed and unimodal, with few items designated as essential by more than 50% of the teachers. As no clear distinction between essential and non-essential items was made by the teachers, the jury designations became the sole criterion for categorizing the test items as representations of essential and non-essential information.

When the distribution of teachers' correct responses to the items designated essential by the professional jury is plotted, a distribution is obtained which is similar to that for teachers' responses to all of the 22 test items.

The resulting distribution is presented in Table 10.
Teacher scores across the items designated essential by
the jury parallel those obtained across the entire set of
items, thus creating a flat, elongated, relatively symmetric distribution of scores. The distribution of teachers'
correct responses to the non-essential items resembles a
more peaked, positively skewed distribution of scores
(see Table 10).

TABLE 10

Distributions of Jury Designations of Essential Items,
Teacher Correct Responses for All 22 Items,
and Correct Responses to Essential and Non-Essential Items

Percent of Designation or Correct Responses	Jury Designations of Essential Items	Correct Responses to All 22 Items	Correct Responses to 13 Essential Items	Correct Responses to 9 Non-Essential Items
75-80		1	1	
70-74		3	2	1
65-69	5	1	1	
60-69		1	1	
55-59	3	1	1	
50-54	5			
45-49			1	
40-44		2	2	
35-39		2	1	
30-34	3	2		2
25-29	1	2	1	1
20-24				
15-19	4	3	1	2
10-14				
5-9	1			
0-4		4	1	3

Table 11 outlines percentages of teachers' correct responses on essential items and non-essential items.

TABLE 11

Percentages of Teachers' Correct Answers
on Essential and Non-Essential Items

Essential Items		Non-Essential Items	
Item Number	Percent Correct	Item Number	Percent Correct
4	42	1	17
*8	16	2	3
9	56	3	19
*12	2	5	72
14	69	6	4
15	45	7	4
16	29	10	26
17	35	11	33
18	40	13	33
19	63		
20	72		
21	80		
22	71		

*These items were dropped due to disproportionate numbers of teachers failing them.

Items 8 and 12 were passed by less than 19% of the teacher respondents. Teachers' responses on these items were distributed across all answer choices, and there were apparently no misleading distractors. As the percentages of teachers able to pass items 8 and 12 were 16% and 2% respectively, their worth as test items was seriously limited and they were dropped from the list of essential items.

A t-test for dependent groups was computed to assess differences in individual teachers' scores between essential and non-essential items. The results indicated that there was a significant difference between performance on

essential and non-essential items. The mean number of essential items answered correctly by the teachers was 6.04, or 54.9% of the items. The mean number of non-essential items answered correctly by the teachers was 2.14, or 23.7% of the items. Examination of the means indicated that the teachers performed better on items characterized as representations of information essential for practitioners to know or apply than on items which were not characterized as representations of such information. Results of the t-test are reported in Table 12.

TABLE 12

t-Test Results of Teacher Performance
on Essential and Non-Essential Items

	Mean in Percent Correct	t-Value	Degrees of Freedom	Probability
Essential Items	54.9			
		22.26	165	.000
Non-Essential Items	23.7			

The professional jury respondents categorized each test item as to whether it: 1) represented information about knowledge of the attributes of grades and marks (knowledge), or 2) represented information about

recommended practices for issuing grades and marks to students (practice). Results of that categorization are reported in Table 13.

TABLE 13

Item Categorization by the Professional Jury

Item Number	Knowledge of Attributes		Knowledge of Practices	
	Number of Responses	Percent of Total	Number of Responses	Percent of Total
1	8	67%	4	33%
2	10	83%	2	17%
3	7	58%	5	42%
4	1	9%	10	91%
5	8	67%	4	33%
6	9	75%	3	25%
7	3	25%	9	75%
8	8	67%	4	33%
9	4	33%	8	67%
10	9	75%	3	25%
11	7	58%	5	42%
12	9	75%	3	25%
13	8	67%	4	33%
14	2	17%	10	83%
15	5	42%	7	58%
16	10	83%	2	17%
17	7	58%	5	42%
18	3	25%	9	75%
19	3	25%	9	75%
20	4	33%	8	67%
21	4	33%	8	67%
22	9	75%	3	25%

The percentage distributions of frequencies with which the jury members designated items as representations of knowledge or practice are presented in Table 14. A clear break in both distributions is noted at the 55% designation rate. Two separate distributions are noted for knowledge and practice designations, one above the 55% rate and one

below the 55% designation rate, thus suggesting two differ-
ent types of designations for knowledge and practice items.
Items designated as representations of knowledge or prac-
tice by 55% or more of the jury members were considered
representations of knowledge or practice. Items designated
as representations of knowledge or practice by less than
55% of the jury members were not considered to be represen-
tations of either knowledge or practice.

TABLE 14

Percentage Distributions of Jury Designations
of Knowledge and Practice Items

Percentages of Designation	Number of Designation Percentages	
	Knowledge	Practice
80-84	2	1
75-79	4	1
70-74		
65-69	4	3
60-64		
55-59	3	1
50-54		
45-49		
40-44	1	3
35-39		
30-34	3	4
25-29	3	4
20-24		
15-19	1	2
10-14		
5-9	1	
0-4		

The following essential items were designated as rep-
resentations of information about knowledge of the attri-
butes of grades and marks: 16, 17, and 22. Essential

items designated as representations of information about
recommended practices for issuing grades and marks to stu-
dents were as follows: 4, 9, 14, 15, 18, 19, 20, and 21.
Table 15 illustrates the percent correct scores the aggre-
gate of teachers earned on essential items designated as
representations of knowledge or practice. Examination of
the mean correct percentages for the two sets of data (45%
for knowledge items and 58% for practice items) indicates
a slight difference in scores, with teachers performing
better on the practice items than on the knowledge items.

TABLE 15

Teacher Score Differences Between Essential Items
Representing Knowledge of the Attributes
of Grades and Marks and Recommended Practices
for Issuing Grades and Marks to Students

Knowledge Items		Practice Items	
Item Number	Percent Correct	Item Number	Percent Correct
16	29	4	42
17	35	9	56
22	71	14	69
		15	45
		18	40
		19	63
		20	72
		21	80
Mean = 45		Mean = 58	

An item correlation matrix which estimates how well
scores on each of the essential items correlates with
scores on other items designated as essential is listed in

Table 16. The correlations were established to estimate how well scores on the items were interrelated. Correlations were computed by using the Phi coefficient.

TABLE 16

Correlation Matrix

	Item 4	Item 9	Item 14	Item 15	Item 16	Item 17
Item 4	1.00					
Item 9	-.08	1.00				
Item 14	-.01	.08	1.00			
Item 15	-.01	.09	.00	1.00		
Item 16	-.03	.05	.03	.12	1.00	
Item 17	.06	.10	.17	.31	.06	1.00
Item 18	-.09	.07	.07	.18	.01	.04
Item 19	.13	.05	.15	.22	-.01	.02
Item 20	-.04	.12	-.08	-.10	-.04	.07
Item 21	-.08	.16	.01	-.06	-.04	-.03
Item 22	-.08	.16	.01	-.06	-.04	-.22

	Item 18	Item 19	Item 20	Item 21	Item 22
Item 18	1.00				
Item 19	.37	1.00			
Item 20	-.07	-.10	1.00		
Item 21	-.14	-.14	.11	1.00	
Item 22	-.05	-.02	.22	.20	1.00

Inspection of the correlation matrix indicates that the inter-item correlations ranged from -.22 to .37. These correlation coefficients indicate that there was little relationship between scores obtained on the items, e.g. correct scores on an item did not correlate with correct scores on other items.

Item-total correlations between scores on the essential items and total scores on the sum of the essential items were computed to assess the extent of agreement between individual items and the aggregate scores. Coefficients for item correlations were computed by using the point-biserial formula and are listed in Table 17.

TABLE 17

Item-Total Score Correlations
for Essential Items

Item Number	Item-Total Correlation
4	-.08
9	.22
14	.12
15	.23
16	.05
17	.17
18	.12
19	.21
20	.02
21	-.01
22	.02

SUMMARY

In Chapter IV, data collected during development of the test items has been presented. The following statements summarize the findings outlined in this chapter.

1. The modal correct score for the Nebraska teachers completing the 22 test items was 10 correct, and scores ranged from 1 to 14 correct.

2. Most of the test items were failed by 50% or more of the teacher respondents. Only

items 5, 9, 14, 19, 20, 21, and 22 were
answered correctly by 50% or more of the
teachers; however, this was more than
half of the items designated essential
by the jury.

3. Pearson product-moment correlation coeffi-
cients computed to assess the degree of re-
lationship between jury designations of es-
sential items, teacher designations of es-
sential items, and teachers' correct re-
sponses indicated that a mild degree of re-
lationship was found between jury designa-
tions of essential items and teachers' cor-
rect responses. The correlation coefficient
for this relationship was .56, and there was
less than a 5% probability that this rela-
tionship was due to chance.

4. The professional jury respondents agreed
that the following test items represented
information which was essential for practi-
tioners to either know or apply to implement
recommended grading and marking systems:
4, 8, 9, 12, 14, 15, 16, 17, 18, 19, 20, 21,
and 22.

5. Items 8 and 12 were dropped from the list of
essential items as teachers answered them
correctly only 16% and 2% of the time, re-
spectively, and a final pool of 11 essen-
tial items was derived.

6. Teachers scored significantly better on es-
sential items than on non-essential items.
The mean correct performance of teachers on
the 11 essential items was 6.04, or 54.9%,
and the mean correct performance on the 9
non-essential items was 2.14, or 23.7%.

7. Professional jury respondents designated the
following essential items as representations
of information about knowledge of the attri-
butes of grades and marks: 16, 17, and 22.

8. Professional jury respondents designated the
following essential items as representations
of information about recommended practices
for issuing grades and marks to students:
4, 9, 14, 15, 18, 19, 20, and 21.

9. Teachers performed only slightly better on essential practice items than on essential knowledge items.

10. Scores on individual essential items, when correlated with scores on other essential items, ranged from -.22 to .37. Performance on any single essential item was generally independent from performance on other items, and there was no significant relationship between scores on any of the essential items.

11. Scores on individual essential items, when correlated with total scores on the aggregate of essential items, ranged from -.08 to .23, indicating little agreement between individual item scores and total scores on essential items.

CHAPTER V

SUMMARY, CONCLUSIONS, DISCUSSION,
AND RECOMMENDATIONS

INTRODUCTION

This last chapter is divided into four sections.
First, a summary of the literature and conclusions gener-
ated from that review are delineated. Second, a summary of
the study and a list of findings are outlined. Third, con-
clusions based upon the findings and related discussion are
presented. Last, implications of the conclusions from the
literature review and the study, as well as recommendations
for practice and further research, are listed.

SUMMARY AND CONCLUSIONS
FROM THE LITERATURE REVIEW

There is consensus among members of the educational
community that there is a need for some form of overt
structure to monitor, chart, and report the academic prog-
ress of students systematically. Grading and marking sys-
tems have traditionally provided the structure to fulfill
this purpose.

Research findings suggest that grades and marks have
served as means to communicate the status of student aca-
demic achievement and have provided information which has
assisted in making decisions about student promotion,

admission to special classes and programs, the presentation
of academic honors, the determination of student eligibil-
ity for scholastic and extracurricular activities, and for
determining future college, vocational, or career plans.
The findings from the research also suggest that there are
numerous problems associated with the application of grad-
ing and marking systems, and the quality of decisions which
can be made with information generated from grading and
marking systems is questioned.

Letter symbols are probably the means most widely
used to express grades and marks. Other forms for their ex-
pression have been observed to lesser degrees, but use of
other forms has varied across school districts. The exact
number of forms for reporting grades and marks is not pres-
ently known, but may be so great as to defy generalization.

The weight of the evidence from research findings indi-
cates that grades and marks should reflect students' aca-
demic achievement, nothing more, and nothing less. None-
theless, many grading and marking systems utilize informa-
tion for evaluating a wide range of student behaviors. In
these systems, evaluations of behavior, attitude, attendance,
and effort are averaged with measures of academic achieve-
ment to report grades and marks. Such practices yield con-
taminated measures of academic achievement which are in-
fluenced by these other factors to some unknown degree.
The resulting measures represent a nebulous compilation of

elements; their meaning and comparability are compromised, and confusion over their interpretation is a probable out-come.

Many of the studies which have assessed the reliability and validity of grades and marks are based on rational or philosophical information, and there has not been sufficient research-based information to adequately document levels of reliability and validity for grades and marks. Conclusions from research related to the reliability and validity of grades and marks outlined in the following paragraphs are based on only a few research-based findings, which have been bolstered by rational or philosophical reports promulgated by authors with knowledge of grading and marking.

Grades have demonstrated little reliability across students. Teachers implement a wide variety of systems for determining and reporting student progress, and grades reported by these systems have demonstrated little comparability. Nonetheless, there are some research findings which suggest that individual student marks demonstrate stability from year to year, thus suggesting that marks possess some degree of reliability. Individual student grades which comprise these marks may possess some degree of stability also, but researchers have not assessed the reliability of grades for individual students.

Research data do not indicate that marks are stable

measures across groups of students. Therefore, marks can be viewed as estimates of the academic performance stability of individual students. They should not, however, be viewed as estimates of the academic performance stability of groups of students, as research data do not support such comparisons.

Researchers have yet to demonstrate that grades possess any degree of validity; there have been no studies located which address the validity of grades. To possess useful validity, grades should do the following: predict future performance accurately, measure academic achievement accurately, and match predictions of achievement obtained from other sources. Researchers have yet to demonstrate statistically that grades can perform any of these functions. Marks, which are comprised of individual grades, have demonstrated reasonable representations of predictive and concurrent validity.

Marks possess a moderately high capacity to predict the future educational performance of individual students; thus, they can demonstrate a moderate degree of predictive validity and probably predict future performance of individual students as well as standardized tests. Marks can possess a moderate degree of concurrent validity, as they correlate well with IQ as a predictor of classroom achievement; they predict achievement as well as standardized tests; and they correlate acceptably with other marks

obtained by a single student in courses of the same subject matter.

Some researchers have suggested that marks can demonstrate reasonable representations of content validity, and detailed methods for appropriately deriving them have been offered; however, these methods involve detailed statistical transformations, and the ability of practitioners to apply such techniques appropriately is suspect. In reality, marks derived by practitioners probably do not always possess a high degree of content validity, as content validity is compromised when statistical rules for deriving marks are not followed.

Decisions based on grading and marking systems have been subjected to litigation in the past, and there is no reason to predict that this will not continue in the future. Results of pertinent litigation suggest the following implications. First, school districts must develop and apply a reasonable, written process to make decisions based on grades or marks. Second, schools must assure that grades and marks demonstrate reliability and validity, as they may be used to facilitate decisions about students which convey serious and far-reaching consequences. Third, grades and marks should remain prominent as standards of academic achievement. Schools should not hesitate to use grading and marking systems to enforce academic standards. Last, grades and marks should generally not be reduced as

sanctions if they create penalties which seriously diminish achievement; nonetheless, they can be legitimately reduced for failure to achieve the academic standards promulgated by schools, or as sanctions for behavior which bears a clear relationship to relevant educational standards. In spite of their potential legality, though, grade or mark reductions do not represent good practice, as they carry little weight as punishers.

The effectiveness of grade and mark reductions for changing student behavior is questioned on the following grounds. First, for grades and marks to serve as punishments, they must become meaningful symbols to students. Grades and marks are probably not meaningful symbols to all students. Withholding a high grade or awarding a low grade or mark as a punishment for failure to achieve certain academic standards will probably have little impact in eliciting behavioral changes for all students.

Second, grade or mark reductions as sanctions for student misbehavior violate the standard that punishment should be related to the transgression. A grade or mark reduction represents an academic penalty for a behavioral transgression.

Third, the minor types of grade (or mark) reductions which have been approved by the courts probably carry little weight as punishments. These are likely to have minimal consequences, and punishments with such slight

consequences have been found to be ineffective in changing behavior.

SUMMARY OF THE STUDY

The Problem

Numerous research studies from the 1950's through the early 1980's have specified problems regarding the application of grading and marking systems in public K-12 schools. Stanley and Hopkins (1972) have theorized that problems associated with grades derived from such systems can be subsumed under four categories of concern: 1) grades are inaccurate and not comparable across instructors, departments, or schools, 2) grades focus upon false and inappropriate objectives which have little relationship to important educational objectives, 3) grades are limited in value as a medium of communication between school and home, and 4) grades are responsible for a variety of detrimental side effects. Definitional inconsistencies have plagued terms associated with grading and marking systems. Researchers, as well as practitioners, have demonstrated confusion over this related terminology. Problems which have been associated with grades and grading systems may likely be associated with marks and marking systems as well.

Despite the breadth of problems associated with grading and marking systems, little has been done to resolve, or even to clarify, controversies associated with these

problems. Grading and marking practices used in schools have been questioned on theoretical, as well as practical, grounds (Hedges, 1969; Hills, 1976; Stefanelli, 1981).

Grading and marking systems are hypothesized to consist of the interaction of attributes of four domains. Rational, psychometric, legal, and learning domains interact to comprise a single grading or marking system. The quality of assumptions about, knowledge of, and applications for the four domains determine the legitimacy of the system; inadequacies inherent in attributes or applications within any domain can invalidate the worth of the entire system.

Persons developing and implementing a grading and marking system should be knowledgeable about the assumptions, attributes, and practices inherent in each of the four domains that comprise the system. Failure to acknowledge principles or to accommodate for inadequacies of practice within any domain is likely to result in the implementation of a system that is weak in either reliability, validity, or both.

Although an extensive body of research literature on problems associated with grading and marking systems has been generated and recommendations for appropriate applications have been delineated, the extent to which findings from this research are acknowledged and applied by educational practitioners is suspect. The present study was

undertaken to design a pool of test items to assess the extent to which research findings associated with the attributes of grades and marks and recommended principles for issuing grades and marks to students are acknowledged and understood by practitioners.

The Procedures

The purpose of this study was to create a pool of test items to assess knowledge of principles and practices associated with the rational, psychometric, legal, and learning domains of grading and marking systems, for which reliability estimates and content validity characterisitcs would be established. The test items were created, refined, and finalized by following a specific test development plan. As the plan was implemented, information was collected to assess two research objectives. These were:

1. To ascertain the extent of agreement among experts, with content knowledge congruent with the four domains of grading and marking systems, as to what information is essential for practitioners to either know or apply to implement grading and marking systems that are consistent with recommended practices.

2. To discover the extent of agreement among teachers who regularly apply grading and marking systems as to what information is essential to either know or apply to implement recommended grading and marking systems.

Information to be included in the test items was

chosen by conducting a content analysis of research find-
ings associated with the study of grading and marking sys-
tems. Multiple-choice items were created to reflect this
information, and the test item stems, foils, and answers
were refined by analyzing qualitative data collected from
university professors, a professional jury, and a sample
of Nebraska teachers.

Data regarding the content, nature, and merit of the
test items were gathered from a professional jury which was
composed of experts possessing knowledge congruent with
attributes of the four domains of grading and marking sys-
tems. Data to complete reliability estimates and item
analyses were collected from a sample of 250 Nebraska
teachers.

Qualitative analyses of the information collected on
stem-foil matches, grammar, punctuation, and the overall
organization of the test items were conducted by weighing
input from university professors and professionals includ-
ed in the validation samples. Quantitative analyses to as-
sist with the item analyses and appraisal of the items were
conducted with the assistance of computer programming at
the Nebraska Evaluation and Research Center on the Univer-
sity of Nebraska-Lincoln campus, by using the Statistical
Package for Social Sciences, Version Nine.

The Findings

Findings of the study were:

1. The modal correct score for the Nebraska teachers completing the 22 test items was 10 correct, and scores ranged from 1 to 14 correct.

2. Most of the test items were failed by 50% or more of the teacher respondents. Only items 5, 9, 14, 19, 20, 21, and 22 were answered correctly by 50% or more of the teachers; however, this was more than half of the items designated essential by the jury.

3. Pearson product-moment correlation coefficients computed to assess the degree of relationship between jury designations of essential items, teacher designations of essential items, and teachers' correct responses indicated that a mild degree of relationship was found between jury designations of essential items and teachers' correct responses. The correlation coefficient for this relationship was .56, and there was less than a 5% probability that this relationship was due to chance.

4. The professional jury respondents agreed that the following test items represented information which was essential for practitioners to either know or apply to implement recommended grading and marking systems: 4, 8, 9, 12, 14, 15, 16, 17, 18, 19, 20, 21, and 22.

5. Items 8 and 12 were dropped from the list of essential items as teachers answered them correctly only 16% and 2% of the time, respectively, and a final pool of 11 essential items was derived.

6. Teachers scored significantly better on essential items than on non-essential items. The mean correct performance of teachers on the 11 essential items was 6.04, or 54.9%, and the mean correct performance on the 9 non-essential items was 2.14, or 23.7%.

7. Professional jury respondents designated the following essential items as representations of information about knowledge of the attributes of grades and marks: 16, 17, and 22.

8. Professional jury respondents designated the following essential items as representations of information about recommended practices for issuing grades and marks to students: 4, 9, 14, 15, 18, 19, 20, and 21.

9. Teachers performed only slightly better on essential practice items than on essential knowledge items.

10. Scores on individual essential items, when correlated with scores on other essential items, ranged from -.22 to .37. Performance on any single essential item was generally independent from performance on other items, and there was no significant relationship between scores on any of the essential items.

11. Scores on individual essential items, when correlated with total scores on the aggregate of essential items, ranged from -.08 to .23, indicating little agreement between individual item scores and total scores on essential items.

CONCLUSIONS AND DISCUSSION
BASED UPON FINDINGS FROM THE STUDY

The Nebraska teacher respondents demonstrated difficulties appropriately answering over two-thirds of the 22 test items. Only items 5, 9, 14, 19, 20, 21, and 22 were passed by 50% or more of the teachers. The modal correct score on the aggregate of 22 test items was 10 correct, and scores ranged from 1 to 14 correct. Difficulties the teachers encountered were most likely attributed to either or both of the following: 1) wording problems or terminology inconsistencies inherent in the test items which would

have made them difficult to understand, 2) a lack of familiarity or disagreement with the content of the test items.

Some teachers reported problems defining terms such as reliability and validity, and researchers have documented confusion over definitions for the terms scoring, grading, and marking. All of these terms were used in many of the test items. Problems related to terminology variables could have limited teachers' performance to an unknown degree.

The wording of the test item stems, foils, and answers was scrutinized by the author and three university professors, and wording was sharpened before items were subjected to validation. Thus, it is unlikely that wording problems contributed to adverse test performance to any significant degree. In addition, inspection of the teachers' response choices across all of the 22 items indicates there were few items in which the incorrect answer choices were loaded on one or two positions, thus suggesting that wording problems did not direct respondents to any particular choices.

A mild relationship (r=.56) was found between jury designations of items deemed essential for practitioners to know or apply and teachers' correct responses on all 22 test items. This suggests that teachers may have learned grading and marking procedures the professional jury members have designated as important to know. The teachers' scores correlated better with the jury members' designations

of essential items than with their own designations of essential items.

Inspection of the distribution of teachers' designations of essential items suggested that there was little agreement among Nebraska teachers regarding which information is essential to know or apply to implement recommended grading and marking practices. Teachers may realize certain levels of knowledge about recommended grading and marking practices, but may fail to agree upon the importance of that knowledge.

Professional jury respondents agreed that 13 of the 22 test items represented information essential for practitioners to know or apply to implement recommended grading and marking systems; hence, professionals who possess knowledge congruent with the four domains of grading and marking systems can agree on which bits of information are essential to implement recommended systems. In this study, a clear demarcation was found between items the professional jury perceived as representations of essential information and items perceived to represent non-essential information. No such demarcation was found for teachers' designations of essential items. As the teachers did not clearly indicate agreement on the items which represented information essential to know or apply, the jury designations of essential items were used to select a pool of items for further analysis.

A pool of 13 test items was selected for further anal-
yses which consisted of the items designated by the jury
as representations of essential information. Inspection of
teachers' responses to the essential items indicated that,
for two of the items, the percent of teachers passing them
was much less than chance. After reviewing teachers' re-
sponses across all of the response choices for the two
items, they were dropped from the item pool, as their abil-
ity to discriminate between teachers' levels of knowledge
was unacceptable. Thus, 11 items were left in the item
pool, and these items were subjected to statistical anal-
yses. The distribution of teachers' responses to the 11
essential items was similar to the distribution of re-
sponses across all 22 items, but scores on essential items
averaged higher than those on all 22 items.   .

Teachers scored significantly better on items desig-
nated as representations of essential information than on
items not so designated (non-essential items). This find-
ing supports the correlation reported between the jury's
designations of essential items and teachers' scores across
all 22 test items. Teachers are more likely to have
learned the information the jury designated as essential
than the information which the jury considered non-essen-
tial.

The 11 essential items were broken down into items
which represented knowledge about the attributes of grades

and marks and items which represented practices for issuing grades and marks to students. Professional jury members categorized all items as representations of knowledge or practice, and a clear demarcation was observed between designations of the essential items as representations of knowledge or practice.

Teachers performed slightly better on essential items representing practices for issuing grades and marks than on essential items representing knowledge of the attributes of grades and marks; however, the validity of this finding is suspect. Only three of the essential items were designated as representations of knowledge. The remaining eight were designated as representations of practice. Conclusions based on such a small sample are dubious. Further, the slight differences reflected in teachers' performances on essential knowledge and practice items may not represent any practical differences.

There was little intercorrelation of scores on the essential items which comprised the item pool. This suggested a random pattern of answers. If respondents passed or failed a given item, there was little relationship between performance on that item and performance on any of the other items. Low intercorrelations of the items constrained the potential reliability of the pool of items, as erratic scores on the items inhibited the establishment of a consistent measurement base.

Possible reasons for the random pattern of answers to the test items could be linked to the overall difficulty level of the items. Only 5 of the 11 items were answered correctly by 49% or more of the teachers, and there were no items which were passed by more than 80% of the teacher respondents. Items with high difficulty levels, such as the ones reported here, could quite possibly have elicited a random pattern of responses, thus constricting the degree of intercorrelation between scores, which would constrain reliability.

Other possible reasons for random patterns of answers to the test items could relate to teacher responses based upon attitudes or beliefs, rather than familiarity with the content of the items. Teachers may not have been familiar with the information delineated in the items, or they may have disagreed with the information contained in the items, and may have responded based upon beliefs or attitudes. Responses of that kind could have served to create a random pattern of answers. Teachers could have been familiar with some, but not all, of the information delineated in the items. High scores could have been earned on some items and low scores on others, and no relationship would be observed between scores.

Items included in the final item pool did not discriminate between high and low performers as determined by the total scores across all 11 items. There was little

relationship between scores on individual items and total scores of higher and lower performers. The performance on individual items was not substantially related to a respondent's level of performance across all items. This absence of a positive relationship could also have served to limit any potential reliability of the pool of test items.

Content validity was built into the test items by using the following methods. First, content analysis of research findings facilitated the inclusion of relevant content into the test items. Second, a jury made up of individuals who possessed knowledge congruent with the four domains of grading and marking systems was identified and asked to complete tasks designed to help define and sharpen the content of the test items.

Completion of the tasks outlined here assured the representation of appropriate content in the test items. Content validity was built into the test as it was developed. Unfortunately, though, as the test items did not demonstrate the potential for a high degree of reliability, the impact of their content validity is limited, as low reliability impedes utility as a consistent measurement instrument. In addition, a low response rate from the professional jury members (50% overall, but 20% from school attorneys and 40% from school psychologists) probably constricted content validity to some unknown degree.

The quality of representation of school psychologists

and school attorneys was questioned. More input from professionals in these areas may or may not have changed definitive statements regarding the content of the test items.

The end result of this study was that a set of 13 test items was created which possessed at least a moderate degree of content validity. Eleven of the items were subjected to statistical analyses, and the 11 items, in their present form, did not demonstrate sufficient reliability to serve as consistent measurement tools. Reliability was hampered by weak inter-item correlations and low item-total score correlations.

Several steps should be taken to move from this set of 11 test items to a useful pool of items capable of assessing attributes of the four domains of grading and marking systems. First, the content of the original 13 items which were deemed to represent information essential for practitioners to know or apply should be outlined, and a statement representing the content of each item should be developed.

Second, several collateral items should be designed which reflect each of the 13 statements of content, until a pool of at least 40 items is created. Third, the items should be broken down into domain categories and each item designated as a representation of the rational, psychometric, legal, or learning domain. In cases in which the designation is not obvious, the item should be assigned to all

domains that seem appropriate.

Fourth, a panel of Nebraska experts should be derived
to represent professionals with knowledge congruent with
the four domains of grading and marking systems. The Ne-
braska experts would then be given only the items represent-
ing their own domain of expertise, and would be asked to
assist in refining the content of the items. The Nebraska
experts would be asked to categorize the items as represen-
tations of knowledge or practice and to provide formative
feedback regarding their worth.

Fifth, the new pool of items should be pilot-tested
on a small sample of teachers. Answers from the teachers
could be plotted to check for response position variables
and misleading foils. Appropriate changes could then be
made before the items are subjected to the final phase of
validation.

Last, the items should be re-administered to a large
sample of teachers. Responses from this sample may then be
used to statistically test the reliability of the items,
and the items which possess the best psychometric proper-
ties can then be included in a finalized item pool.

IMPLICATIONS AND RECOMMENDATIONS

Implications of the Study

Implications of the study which are based on the con-
clusions are outlined as follows. Professionals who possess

knowledge congruent with the four domains of grading and marking practices can agree on the kinds of information which should be considered in developing and implementing grading and marking systems. Researchers and practitioners across the United States do not agree on substantive terminology definitions, and there is no single grading and marking system recommended for use in all applications. The systems which are used in schools represent attempts to develop systems responsive to individual school or school district needs, and there is generally little comparability between systems or between the grades and marks which are generated. In addition, research findings document many problems associated with traditional grading and marking systems, yet they continue to be used in schools.

Nebraska teachers probably do not agree on the kinds of information which should be considered in developing and implementing grading and marking systems, and Nebraska teachers likely mirror the confusion of other professional groups on substantive terminology definitions. Nebraska teachers may be unfamiliar with research-based recommendations for developing and implementing grading and marking systems.

Recommendations

Recommendations for practice and further research which are based upon conclusions and implications generated from the study are listed as follows.

1. A common language for substantive terminology related to grading and marking systems should be recognized and used to express relevant information. The following terms should be used only according to the definitions which accompany them.
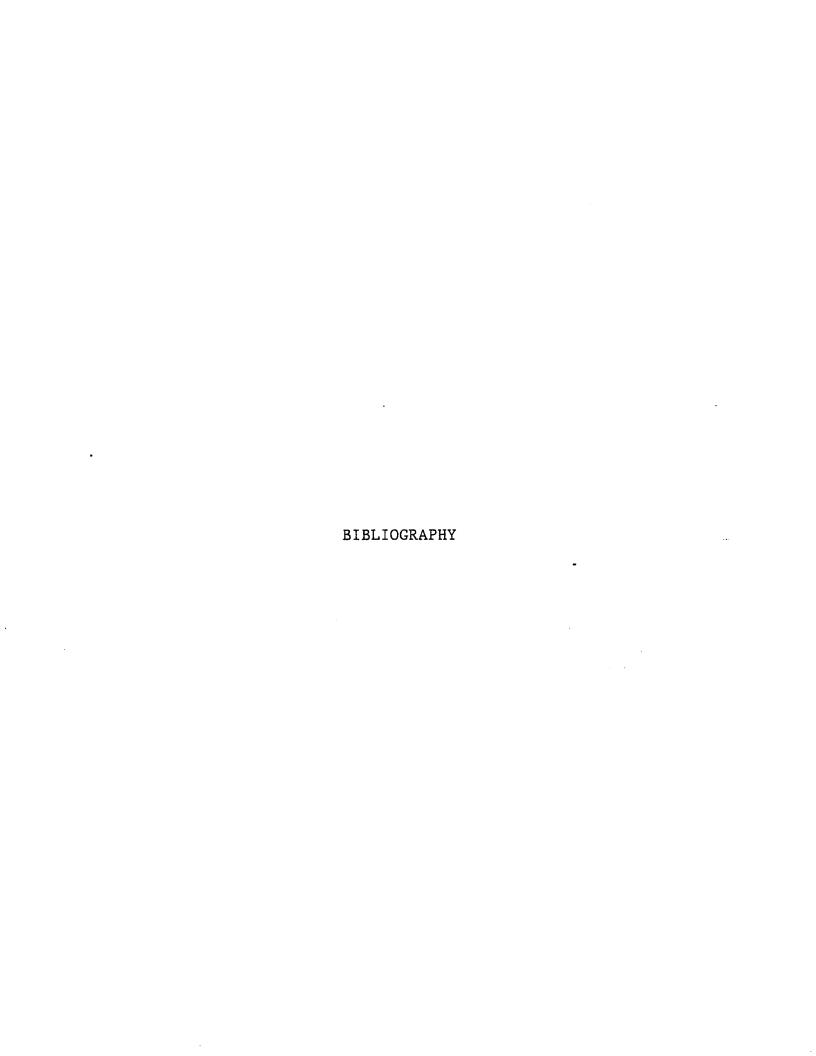
   Scoring - The process of correcting student work and calculating a raw or derived score based on the assessed performance.

   Grading - The attachment of a grade (usually a letter or number symbol) to a student product based upon comparison of the performance score to some standard.

   Marking - The assignment of a composite grade symbol (usually a letter or number) to represent the cumulative achievement of a student as measured on a variety of products.

2. Each school district should develop policies to be implemented by its governing board which specify the composition of its grading and marking system, methods used to derive grades and marks, and procedures used to make decisions based on grades and marks.

3. School districts should develop district-wide grading and marking systems responsive to their needs, as reflected in governing board policies. Such systems should delineate policies for determining and issuing grades and marks, and attributes of the systems should reflect research findings within the rational, psychometric, legal, and learning domains.

4. School districts should conduct research studies to assess the extent of reliability and validity inherent in the grading and marking systems which are implemented.

5. In addition to reliability and validity studies, each school district should formally evaluate the worth of its grading and marking system at least once every three to five years.

6. Preservice training, inservice training, or both should be utilized to familiarize teachers with information contained in research findings regarding the composition, merits, and liabilities of grading and marking systems.

7. Preservice training, inservice training, or both should be utilized to train teachers in appropriate methods for deriving and reporting grades and marks.

8. The pool of 11 test items which was developed in this study should be expanded and re-validated according to the steps which have been outlined in this chapter.

9. Once a pool of reliable and valid test items has been developed, it can be used to identify strengths and weaknesses of knowledge about grading and marking systems; appropriate activities for remediating weaknesses can then be implemented.

BIBLIOGRAPHY

BIBLIOGRAPHY

BOOKS

Ahmann, J. S., and Glock, D. Evaluating Student Growth. (6th Ed.) Boston: Allyn & Bacon, 1981.

Brown, F. G. Measurement and Evaluation. F. E. Peacock, 1971.

Chase, C. I. Measurement for Evaluation. (2nd Ed.) Reading, Mass.: Addison-Wesley Publishing Company, 1978.

Cronbach, L. J. Educational Psychology. (2nd Ed.) New York: Harcourt, Brace, & World, 1963.

Davis, F. B. Educational Measurements and Their Interpretation. Belmont, Calif.: Wadsworth Printing Company, 1964.

Dressel, P. L. Handbook of Academic Evaluation. London: Jossey Bass, 1976.

Ebel, R. L. Measuring Educational Achievement. Englewood Cliffs, N.J.: Prentice-Hall Inc., 1965.

Ebel, R. L. Essentials of Educational Measurement. (3rd Ed.) Englewood Cliffs, N.J.: Prentice-Hall Inc., 1979.

Erickson, R. L., & Wentling, T. L. Measuring Student Growth. Boston: Allyn & Bacon Inc., 1976.

Gatti, R. G., & Gatti, D. F. Encyclopedic Dictionary of School Law. West Nyack, N.Y.: Parker Publishing Company, 1975.

Gay, L. R. Educational Evaluation and Measurement. Columbus, Ohio: Charles E. Merrill Publishers, 1980.

Green, J. Introduction to Measurement and Evaluation. New York: Dodd, Mead, & Company, 1970.

Gronlund, N. E. Measurement and Evaluation in Teaching. (4th Ed.) New York: Macmillan Company, 1981.

Guilford, J. P., & Fruchter, B. Fundamental Statistics in Psychology and Education. (5th Ed.) New York: McGraw-Hill, 1973.

Hedges, W. Evaluation in the Elementary School. New
     York: Holt, Rinehart, & Winston, 1969.

Hilgard, E. R., & Bower, G. H. Theories of Learning. (3rd
     Ed.) New York: Appleton-Century-Crofts, 1966.

Hills, J. Measurement and Evaluation in the Classroom.
     Columbus: Charles E. Merrill, 1976.

Karmel, L., & Karmel, M. Measurement and Evaluation in the
     Schools. New York: Macmillan, 1978.

Klausmeier, H. J., & Goodwin, W. Learning and Human Abil-
     ities. (2nd Ed.) New York: Harper & Row, 1966.

Lavin, D. E. The Prediction of Academic Performance. New
     York: Wiley, 1965.

Lien, A. J., & Lien, H. S. Measurement and Evaluation of
     Learning. (4th Ed.) Dubuque, Iowa: William C. Brown
     Company, 1980.

Lindeman, R. H. Educational Measurement. Glenview, Ill.:
     Scott, Foresman, & Company, 1967.

Lindvall, C. M. Testing and Evaluation: An Introduction.
     New York: Harcourt, Brace, & World Inc., 1961.

McCarthy, M. M., & Gambron, W. A. Public School Law Teach-
     ers' and Students' Rights. Boston: Allyn & Bacon,
     1981.

McIntosh, H. G. (Ed.). Techniques and Problems of Assess-
     ment. London: Edward Arnold Publishing, 1974.

Mehrens, W. A., & Lehmann, I. J. Measurement and Evalua-
     tion in Education and Psychology. New York: Holt,
     Rinehart, & Winston, 1973.

Mitzel, H. E., Best, J. H., & Rabinowitz, W. (Eds.). Ency-
     clopedia of Educational Research. (5th Ed.) New York:
     Macmillan, 1982.

Mussen, P. H., Conger, J. J., & Kagan, J. Child Develop-
     ment and Personality. (2nd Ed.) New York: Harper &
     Row, 1963.

Nolte, M. C. Nolte's School Law Desk Book. West Nyack,
     N.Y.: Parker Publishing Company, 1980.

Nunnally, J. C. Psychometric Theory. New York: McGraw-
     Hill, 1967.

Nunnally, J. C., & Ator, N. A. Educational Measurement and Evaluation. New York: McGraw-Hill, 1972.

Odell, E. W. Marks and Marking Systems. In W. S. Monroe (Ed.), Encyclopedia of Educational Research. (2nd Ed.) New York: Macmillan, 1950, pp. 711-717.

Pavlov, I. P. Conditioned Reflexes. London: Oxford University Press, 1927.

Payne, D. A. The Specification of Learning Outcomes. Waltham, Mass.: Blaisdell Publishing Company, 1968.

Skinner, B. F. The Behavior of Organisms: An Experimental Analysis. New York: Appleton-Century-Crofts, 1938.

Smith, F. M., & Adams, S. Educational Measurement for the Classroom Teacher. New York: Harper & Row, 1972.

Stallings, W. Student Attitudes Towards Grades and Grading Practices. Urbana, Ill.: University of Illinois Press, 1968.

Stanley, J. C., & Hopkins, K. Educational and Psychological Measurement and Evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Terwillinger, J. S. Assigning Grades to Students. Glenview, Ill.: Scott, Foresman, & Company, 1971.

Terwillinger, J. S. Individual differences in the marking practices of secondary school teachers. In Noll, V. H., Scannell, D. P., & Wall, R. P., Introductory Readings in Educational Measurement. New York: Houghton Mifflin, 1972.

Thorndike, E. L. Animal Intelligence. New York: Macmillan, 1911.

Travers, R. M. Essentials of Learning. (4th Ed.) New York: Macmillan Company, 1977.

Valente, W. D. Law in the Schools. Columbus, Ohio: Merrill Publishing Company, 1980.

Walters, G. D., & Grusec, J. E. Punishment. San Fransisco: W. H. Freeman & Company, 1977.

PERIODICALS

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

Bendig, A. W. The reliability of letter grades. Educational and Psychological Measurement, 1953, 13, 311-321.

Bippus, S. A new look at progress reports. NASSP Bulletin, May, 1981, 65, 109-111.

Bloom, R., & Bourdon, L. Types and frequencies of teachers' written instructional feedback. Journal of Educational Research, Sept./Oct., 1980, 74, 13-15.

Branwaite, A., Trueman, M., & Berrisford, T. Unreliability of marking: further evidence and a possible explanation. Educational Review, 1981, 33, 41-46.

Broadfoot, P. Communication in the classroom: a study of the role of assessment in motivation. Educational Review, Nov., 1979, 31, 3-10.

Cullen, F. T., Cullen, D. B., Hayhow, V. L., & Plouffe, J. T. The effects of the use of grades as an incentive. Journal of Educational Research, 1975, Vol 68 No. 7, 277-279.

Hills, J. Predictions of grades for all public colleges of a state. Journal of Educational Measurement, 1964, 1, 155-159.

Hull, R. Fairness in grading - perceptions of junior high school students. Clearing House, May, 1980, 53, 340-343.

Hurlock, E. B. Child Development. (6th Ed.) New York: McGraw-Hill, 1978.

Jennings, E. K. A due process anomaly. Nebraska Law Review, 1979, Vol 58 No. 2, 519 ff.

Lesser, G. S., Davis, E. B., & Nahemar, L. G. The identification of gifted elementary school children with exceptional scientific talent. Educational and Psychological Measurement, 1962, 22, 349-363.

Marshall, K. Making learning meaningful. Learning, Spring, 1981, 10, 57.

Newman, B., & Wise, R. I. The responsibility of grading.
    Educational Leadership, Jan., 1975, 32, (4), 253-256.

Russell, I. L., & Thalman, W. A. Personality: does it
    influence teachers' marks? Journal of Educational
    Research, 1955, 48, 561-564.

Salvia, J. Attractiveness and school achievement. Jour-
    nal of School Psychology, Spring, 1977, 15, 60-67.

Slavin, R. E. Effects of individual learning expectations
    on student achievement. Journal of Educational Psy-
    chology, 1980, 72, 520-524.

Stefanelli, J. Student success. NASSP Bulletin, Feb.,
    1981, 65, 109-111.

Yarborough, B., & Johnson, R. Research that questions the
    traditional marking system. Phi Delta Kappan, April,
    1980, 61, 327-528.

OTHER SOURCES

Anderson, K. E. Grading: current practices and problems.
    Unpublished Manuscript, 1982.

Anderson v. Board of Education, U.S. District Court Opinion.
    (1977).

Barnard v. Shelbourne, 216 Mass. 19 102 N.E. 1095 (1913).

Board of Curators of the University of Missouri et. al. v.
    Horowitz, 435 U.S. 78 (1978).

Board of Education of the Hendrick Hudson Central School
    District v. Rowley, 102 U.S. 3034 (1982).

Connelly v. University of Vermont, 244 F. Supp. 156 (1965).

Debra P. v. Turlington, 644 F. 2d 397 (1981).

Dorsey v. Bale, 521 S.W. 2d 76 (1975).

Foley v. Benedict, 122 Tex. 193 55 S.W. 2d 805 (1932).

Fricke, B. G. Report to the Faculty: Grading, Testing,
    Standards, and all That. Ann Arbor: University of
    Michigan Evaluation and Examination Office, 1975.

Gardner, M. R.   Punishment in the Schools.   Unpublished
    Manuscript, 1983.

Gaspar v. Bruton, 513 F. 2d 843 (1975).

Goss v. Lopez, 419 U.S. 565, 573 (1975).

Hamer v. Board of Education of Township High School, 383
    N.E. 2d 231 (1978).

Johnson v. Shineman.   National Organization on Legal Prob-
    lems of Education.   Nolpe Notes, Aug., 1982, Vol. 17
    No. 8, p. 5.

Knight v. Board of Education, 348 N.E. 2d 299 (1976).

Mahavangsen v. Hall, 529 F. 2d 448 (1976).

Miles, W. R.   Simformation 5:   Reporting Student Progress.
    Report from the Project on Home-School-Community Re-
    lations, Wisconsin University, Madison.   Research and
    Development Center for Cognitive Learning, Nov., 1977.

Minorics v. Board of Education of Phillipsburg.   Decision
    of the New Jersey Commissioner of Education, March 24,
    1972.

Mustel v. Rose, 282 Ala. 358, 367 211 So. 2d 489, 498
    (1968).

National Association of Secondary School Principals.   Grade
    reduction, academic dismissal and the courts.   A Legal
    Memorandum, Oct., 1971.

Navato v. Sletten, 415 F. Supp. 312 (1976).

Reporting Pupil Progress to Parents 1966 & 1971.   (NEA Re-
    search Memo 1972-10).   Washington D.C.:   the Associa-
    tion, Sept., 1972, 1.

Rinne, H.   Grading and growth: answer to an editorial.
    Educational Leadership, Jan., 1975, 32 (4), 247-249.

Robinson, J. T.   A critical look at grading and evaluation
    practices.   Unpublished manuscript, 1978.

Sandlin v. Johnson, 643 F. 2d. 1027 (1981).

State of Connecticut v. Teal, Supreme Court Docket No. 80-
    2147 (1982).

The National Commission on Excellence in Education. A Nation at Risk: The Importance of Educational Reforms. Washington D.C. U.S. Government Printing Office, 1983.

Warner, T. A. Alternatives to ABC Grades. Paper presented at the 62nd Annual Meeting of the National Association of Secondary School Principals, Anaheim, Feb., 1978.

Wermuth v. Bernstein & Board of Education of the Township of Livingston, Essex County. Decision of the New Jersey Commissioner of Education, 1965.

APPENDIX A

Data Collection Materials

1208 Westridge Drive
Blair, Nebraska 68008
(402) 426-9675
April 9, 1984

I am a doctoral candidate in Educational Administration at
the University of Nebraska-Lincoln. As a part of my dis-
sertation, I am constructing a test to assess knowledge of
the attributes of grades and marks and awareness of recom-
mended principles for issuing them to students.

As one step in the process of validating the test, I am
assembling a panel of experts to critique the test items.
Would you assist me in this process? Please do so by
providing me with the names, and addresses if possible, of
at least five colleagues I can contact whom you perceive to
be the most renowned experts in School Psychology. Please
list their names and addresses on the attached sheet and
return it to me in the enclosed, self-addressed, stamped
envelope as soon as possible, but not later than April 20,
1984.

Thank you for your cooperation. If you have any questions,
please feel free to write or call me at your convenience.
My home address and phone number are listed above. My work
phone number is (402) 426-2610.

                              Sincerely,

                              Kenneth E. Anderson

Code __126__

NAME	ADDRESS
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____

1208 Westridge Drive
Blair, Nebraska 68008
(402) 426-9675
April 23, 1984

Two weeks ago, on April 9, 1984, I sent you a letter asking for
your assistance in assembling a panel of experts to help validate
a test for assessing knowledge of attributes of grades and marks.
In that letter I asked you to provide me with the names, and ad-
dreses if possible, of at least five colleagues whom you per-
ceive to be the most renowned experts in Educational Psychology.

If you have already completed a list of names and mailed it to
me, please disregard this letter. If however, you have not yet
compiled a list for this project, would you please consider
doing so at this time?

I am enclosing a duplicate copy of the listing sheet provided
in your first mailing. Please complete it and return it to me
as soon as possible, but not later than May 7, 1984.

Thank you for your cooperation. If you have any questions,
please feel free to write or call me at your convenience. My
home address and phone number are listed above. My work phone
number is (402) 426-2610.

Sincerely,

Kenneth E. Anderson

NAME

ADDRESS

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

NAME

ADDRESS

811 Ogden
Oxford, Nebraska 68967
(308) 824-3397
September 17, 1984

I am a doctoral candidate in Educational Administration at
the University of Nebraska-Lincoln. As a part of my dis-
sertation, I am constructing a test to assess two things:
knowledge of the attributes of grades and marks, and aware-
ness of recommended principles for issuing grades and marks
to students. Items that comprise the test have been derived
from a content analysis of research findings applicable to
grades and marks.

Your name has been submitted to me through a peer nomination
process as that of a national expert in the field of School
Law. As one step in validating the test, I am asking experts
from several disciplines to analyze the test items from their
own professional frames of reference, and provide feedback
and further input designed to sharpen the test instrument.
Once the test has been refined, I plan to administer it to
educational practitioners in the state of Nebraska. The re-
sults from this assessment may then be utilized to identify
potential training needs.

Based on your professional knowledge, would you assist me
in the process of validating the test? Please follow the
directions on the attached page, complete the item work-
sheet, and return all pages to me in the self-addressed,
stamped envelope as soon as possible, but not later than
October 1, 1984. If I do not hear from you by the 1st, I
will be sending a follow-up letter which you should receive
around October 3, 1984.

Thank you in advance for your assistance. If you have any
questions, please feel free to call or write me at your
convenience. My home address and phone number are listed
above. My work phone number is (308) 824-3209.

Sincerely,

Kenneth E. Anderson

ITEM WORKSHEET

Items to be included in the finalized test instrument are listed on the following pages. The letter corresponding to the correct answer is circled for each item.

I. Beginning with question #1, sort the items on the following pages into two categories: items that measure knowledge of the attributes of grades and marks, and items that measure knowledge of recommended principles for issuing grades and marks to students. List the number of each item under the appropriate heading below. Place each item in only one category.

Knowledge of the attributes of grades and marks

Knowledge of recommended principles for issuing grades and marks to students

_____   _____

_____   _____

_____   _____

_____   _____

II. List the numbers of items that represent pieces of information you deem essential for educators to either understand or apply to insure good grading and marking practices.

_____

_____

_____

_____

III. Are there any concepts or issues not included in the test instrument that, in your opinion, should be included? If so, please list them below.

_____

_____

_____

_____

_____

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may or may not be congruent with the practices you use).

_____ 1. The process of marking:

    1 = assigns some symbol to represent the achievement of a student on a single product.
    ② = assigns some symbol to represent the cumulative achievement of a student on an articulated set of products.
    3 = assigns a raw or derived score to represent the achievement of a student on a given product.
    4 = all of the above.
    5 = none of the above.

_____ 2. Marks are valid as:

    ① = predictors of the future performance of individual students.
    2 = predictors of the future performance of groups of students.
    3 = a means for accurately communicating with parents about individual student performances.
    4 = statistical transformations for raw data gathered about student performances.
    5 = all of the above.

_____ 3. Low grades and marks issued to students will probably:

    1 = cause most students to try harder.
    2 = affect secondary students more than primary level students.
    3 = insure tough academic standards.
    ④ = carry few negative consequences for most students.

_____ 4. Providing students multiple opportunities or "trials" to demonstrate attainment of instructional objectives:

    ① = is a good idea.
    2 = should be used with all students in a course which is criterion-referenced.
    3 = should be used only with students who have experienced difficulty or are in danger of failing.
    4 = reduces the reliability and validity of grades.
    5 = none of the above.

_____ 5. All other things being equal, which of the following marking practices will yield the most reliable marks?

    1 = assigning numbers one through five.
    2 = assigning marks of A, B, C, D, or F.
    3 = assigning only pass or fail marks.
    ④ = using percentile ranks.

_____ 6. Good grades and marks will probably:

    1 = increase the motivation level of most students who receive them.
    ② = have little impact on most students.
    3 = affect primary level students more than secondary level students.
    4 = affect secondary level students more than primary level students.

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may or may not be contruent with the practices you use).

_____ 7. Decisions made in assigning grades and marks:

①= may be subjected to judicial review.
2 = will not be subjected to judicial review because the courts have traditionally left educational decisions to professional educators.
3 = should be bound by the same due process concerns as decisions to discipline students for misbehavior.
4 = should be reviewed by administrators.

_____ 8. The reliability of grades can be improved by:

1 = including non-cognitive factors in the assessment of student performance.
2 = emphasizing evaluation rather than assessment.
3 = emphasizing gain rather than terminal behavior.
4 = using a two-category system, such as pass-fail or satisfactory-unsatisfactory.
⑤= increasing the number of grade categories.

_____ 9. To develop good grading and marking systems, practicing educators should:

1 = adhere to a standardized system throughout an entire school district.
2 = be familiar with systematic measurement techniques.
3 = develop consistent means for scoring student work and assigning grades and marks.
4 = teach and assess the same instructional content in equivalent courses of study.
⑤= all of the above.

_____ 10. Grading is defined as the process of:

1 = correcting student work.
2 = assigning cumulative grade symbols to report cards.
③= attaching a symbol to a student product.
4 = all of the above.
5 = none of the above.

_____ 11. Giving "extra credit" assignments and opportunities for a student to improve a grade:

1 = is a good idea.
2 = should be used only for students who have experienced difficulty competing with classmates in attaining levels of academic achievement.
3 = should be used only with students who are in danger of failing a course.
④= reduces the reliability and validity of the grade.

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may or may not be congruent with the practices you use).

_____ 12. When combining scores, if a teacher wishes to give equal weight to several scores, best results will be obtained:

1 = if percentile scores are used.
2 = if percentage scores are used.
3 = if raw scores are used.
④ = if standard scores are used.
5 = any of the above—it makes no difference.

_____ 13. Scoring is defined as the process of:

① = correcting student work and calculating the accuracy of performance.
2 = statistically combining several measures of a student's performance to devise a cumulative score.
3 = calculating the distribution of student performances on a given task.
4 = correcting student work and assigning grade symbols to represent levels of performance.
5 = none of the above.

_____ 14. Reducing grades or marks as a punishment for a student's misbehavior in school:

1 = should be done only in required courses of study.
2 = should only be done as a "last resort" when all other means of discipline have failed.
③ = should generally not be done.
4 = should be done as a matter of routine practice.
5 = none of the above.

_____ 15. If a single grade is given, it should represent:

① = academic achievement.
2 = attendance.
3 = behavior.
4 = effort.
5 = all of the above.

_____ 16. How do grades determined by level of improvement (growth) compare with grades determined by level of achievement (status)?

1 = Growth grades are more reliable and more valid as measures of proficiency.
2 = Growth grades are more reliable but less valid as measures of proficiency.
3 = Growth grades are less reliable but more valid as measures of proficiency.
④ = Growth grades are less reliable and less valid as measures of proficiency.

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may or may not be congruent with the practices you use).

_____ 17. Marks in a course should be assigned and interpreted as:

1 = a combination of absolute measures of achievement, relative measures of achievement, and evaluations of students.
② = measures of the level of student achievement.
3 = measures of growth.
4 = evaluations of students.
5 = none of the above.

_____ 18. Reducing grades as a punishment for tardiness in completing an assignment:

1 = is a good idea.
2 = should be done as a matter of policy.
3 = should be used only with those students who are habitually late in submitting assignments.
④ = should not be done.

_____ 19. Reducing a student's grade or mark as a sanction for truancy should:

1 = be done.
② = generally not be done.
3 = be done only in required courses of study.
4 = be done only in non-required courses of study.

_____ 20. In a comprehensive system of marking, separate marks should be reported for:

1 = academic achievement and attendance.
2 = academic achievement and effort. .
3 = academic achievement and behavior.
4 = attendance and effort and achievement.
⑤ = academic achievement, attendance, effort, and behavior.

_____ 21. Grades and marks should be used to:

1 = communicate the status of student academic achievement.
2 = assist in making decisions about retention or promotion.
3 = assist in determining the eligibility of students for participation in the activities curriculum at school.
4 = monitor the academic progress of students.
⑤ = all of the above.

_____ 22. Which set of characteristics should be considered most important in designing a grading and marking system?

1 = content validity.
2 = reliability.
3 = "face" validity among teachers.
4 = "face" validity among students and parents.
⑤ = all of the above.

811 Ogden
Oxford, Nebraska 68967
(308) 824-3397
October 1, 1984

Two weeks ago, on September 17, 1984, I sent you a letter asking for your assistance in validating a test on grades and marks. I have not yet received your response.

If you have already completed the item worksheet that was enclosed with my original letter and placed it in the mail, please disregard this letter. If however, you have not yet completed the item worksheet, would you please consider doing so at this time? Your responses are extremely important.

I am enclosing a duplicate copy of the item worksheet and the test items. Please follow the directions on the item worksheet and return all attached pages to me in the self-addressed, stamped envelope as soon as possible, but not later than October 15, 1984.

Thank you again for your assistance. If you have any questions, please feel free to call or write me at your convenience. My home address and phone number are listed above. My work phone number is (308) 824-3209.

Sincerely,

Kenneth E. Anderson

ITEM WORKSHEET

Items to be included in the finalized test instrument are listed on the following pages. The letter corresponding to the correct answer is circled for each item.

I. Beginning with question #1, sort the items on the following pages into two categories: items that measure knowledge of the attributes of grades and marks, and items that measure knowledge of recommended principles for issuing grades and marks to students. List the number of each item under the appropriate heading below. Place each item in <u>only one</u> category.

<u>Knowledge of the attributes of grades and marks</u>

<u>Knowledge of recommended principles for issuing grades and marks to students</u>

_____   _____

_____   _____

_____   _____

_____   _____

II. List the numbers of items that represent pieces of information you deem essential for educators to either understand or apply to insure good grading and marking practices.

_____

_____

_____

_____

III. Are there any concepts or issues not included in the test instrument that, in your opinion, should be included? If so, please list them below.

_____

_____

_____

_____

_____

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES
AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may
or may not be congruent with the practices you use).

_____ 1. The process of marking:

    1 = assigns some symbol to represent the achievement of a
        student on a single product.
    ② = assigns some symbol to represent the cumulative achieve-
        ment of a student on an articulated set of products.
    3 = assigns a raw or derived score to represent the achieve-
        ment of a student on a given product.
    4 = all of the above.
    5 = none of the above.

_____ 2. Marks are valid as:

    ① = predictors of the future performance of individual students.
    2 = predictors of the future performance of groups of students.
    3 = a means for accurately communicating with parents about
        individual student performances.
    4 = statistical transformations for raw data gathered about
        student performances.
    5 = all of the above.

_____ 3. Low grades and marks issued to students will probably:

    1 = cause most students to try harder.
    2 = affect secondary students more than primary level students.
    3 = insure tough academic standards.
    ④ = carry few negative consequences for most students.

_____ 4. Providing students multiple opportunities or "trials" to
      demonstrate attainment of instructional objectives:

    ① = is a good idea.
    2 = should be used with all students in a course which is
        criterion-referenced.
    3 = should be used only with students who have experienced
        difficulty or are in danger of failing.
    4 = reduces the reliability and validity of grades.
    5 = none of the above.

_____ 5. All other things being equal, which of the following marking
      practices will yield the most reliable marks?

    1 = assigning numbers one through five.
    2 = assigning marks of A, B, C, D, or F.
    3 = assigning only pass or fail marks.
    ④ = using percentile ranks.

_____ 6. Good grades and marks will probably:

    1 = increase the motivation level of most students who receive
        them.
    ② = have little impact on most students.
    3 = affect primary level students more than secondary level
        students.
    4 = affect secondary level students more than primary level
        students.

_____ 7. Decisions made in assigning grades and marks:

(1) = may be subjected to judicial review.
2 = will not be subjected to judicial review because the courts
    have traditionally left educational decisions to profes-
    sional educators.
3 = should be bound by the same due process concerns as
    decisions to discipline students for misbehavior.
4 = should be reviewed by administrators.

_____ 8. The reliability of grades can be improved by:

1 = including non-cognitive factors in the assessment of
    student performance.
2 = emphasizing evaluation rather than assessment.
3 = emphasizing gain rather than terminal behavior.
4 = using a two-category system, such as pass-fail or satis-
    factory-unsatisfactory.
(5) = increasing the number of grade categories.

_____ 9. To develop good grading and marking systems, practicing
    educators should:

1 = adhere to a standardized system throughout an entire
    school district.
2 = be familiar with systematic measurement techniques.
3 = develop consistent means for scoring student work and
    assigning grades and marks.
4 = teach and assess the same instructional content in
    equivalent courses of study.
(5) = all of the above.

_____ 10. Grading is defined as the process of:

1 = correcting student work.
2 = assigning cumulative grade symbols to report cards.
(3) = attaching a symbol to a student product.
4 = all of the above.
5 = none of the above.

_____ 11. Giving "extra credit" assignments and opportunities for a
    student to improve a grade:

1 = is a good idea.
2 = should be used only for students who have experienced
    difficulty competing with classmates in attaining levels
    of academic achievement.
3 = should be used only with students who are in danger of
    failing a course.
(4) = reduces the reliability and validity of the grade.

_____ 12. When combining scores, if a teacher wishes to give equal weight to several scores, best results will be obtained:

1 = if percentile scores are used.
2 = if percentage scores are used.
3 = if raw scores are used.
④ = if standard scores are used.
5 = any of the above—it makes no difference.

_____ 13. Scoring is defined as the process of:

① = correcting student work and calculating the accuracy of performance.
2 = statistically combining several measures of a student's performance to devise a cumulative score.
3 = calculating the distribution of student performances on a given task.
4 = correcting student work and assigning grade symbols to represent levels of performance.
5 = none of the above.

_____ 14. Reducing grades or marks as a punishment for a student's misbehavior in school:

1 = should be done only in required courses of study.
2 = should only be done as a "last resort" when all other means of discipline have failed.
③ = should generally not be done.
4 = should be done as a matter of routine practice.
5 = none of the above.

_____ 15. If a single grade is given, it should represent:

① = academic achievement.
2 = attendance.
3 = behavior.
4 = effort.
5 = all of the above.

_____ 16. How do grades determined by level of improvement (growth) compare with grades determined by level of achievement (status)?

1 = Growth grades are more reliable and more valid as measures of proficiency.
2 = Growth grades are more reliable but less valid as measures of proficiency.
3 = Growth grades are less reliable but more valid as measures of proficiency.
④ = Growth grades are less reliable and less valid as measures of proficiency.

_____ 17. Marks in a course should be assigned and interpreted as:

1 = a combination of absolute measures of achievement, relative
    measures of achievement, and evaluations of students.
② = measures of the level of student achievement.
3 = measures of growth.
4 = evaluations of students.
5 = none of the above.

_____ 18. Reducing grades as a punishment for tardiness in completing
an assignment:

1 = is a good idea.
2 = should be done as a matter of policy.
3 = should be used only with those students who are habitually
    late in submitting assignments.
④ = should not be done.

_____ 19. Reducing a student's grade or mark as a sanction for truancy
should:

1 = be done.
② = generally not be done.
3 = be done only in required courses of study.
4 = be done only in non-required courses of study.

_____ 20. In a comprehensive system of marking, separate marks should
be reported for:

1 = academic achievement and attendance.
2 = academic achievement and effort.
3 = academic achievement and behavior.
4 = attendance and effort and achievement.
⑤ = academic achievement, attendance, effort, and behavior.

_____ 21. Grades and marks should be used to:

1 = communicate the status of student academic achievement.
2 = assist in making decisions about retention or promotion.
3 = assist in determining the eligibility of students for
    participation in the activities curriculum at school.
4 = monitor the academic progress of students.
⑤ = all of the above.

_____ 22. Which set of characteristics should be considered most impor-
tant in designing a grading and marking system?

1 = content validity.
2 = reliability.
3 = "face" validity among teachers.
4 = "face" validity among students and parents.
⑤ = all of the above.

811 Ogden
Oxford, Nebraska 68967
(308) 824-3397
November 6, 1984

I am a doctoral candidate in Educational Administration at
the University of Nebraska-Lincoln. As a part of my disser-
tation, I am constructing a test to assess two things:
practitioner knowledge of the attributes of grades and marks,
and awareness of recommended principles for issuing grades
and marks to students. Items that comprise the test instru-
ment have been derived from a content analysis of research
findings applicable to grades and marks.

My final step in validating the test instrument involves
collecting input from educational practitioners. You have
been selected for inclusion in the validation study by means
of a random sampling of Nebraska educators. A coding pro-
cedure has been utilized solely for my benefit in collecting
responses; however, your responses will be held in strict
confidence. Your responses will not be analyzed or reported
individually, as the validation process I am employing uti-
lizes group response data for each test item.

I realize your time is extremely valuable, so I have attempted
to streamline the validation process as much as possible.
Please take a few minutes to complete the following steps.
First, read the twenty-two questions and answers on the fol-
lowing pages and write the number of the correct answer for
each question in the appropriate space. Second, complete the
item worksheet attached to the test instrument. Last, return
the test instrument and item worksheet to me in the enclosed
self-addressed, stamped envelope as soon as possible, but
not later than November 21, 1984.

Thank you for your assistance. If you have any questions,
please feel free to call or write to me at my home address
listed above.

Sincerely,

Kenneth E. Anderson

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES
AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may
or may not be congruent with the practices you use).

_____ 1. The process of marking:

1 = assigns some symbol to represent the achievement of a
  student on a single product.
2 = assigns some symbol to represent the cumulative achieve-
  ment of a student on an articulated set of products.
3 = assigns a raw or derived score to represent the achieve-
  ment of a student on a given product.
4 = all of the above.
5 = none of the above.

_____ 2. Marks are valid as:

1 = predictors of the future performance of individual students.
2 = predictors of the future performance of groups of students.
3 = a means for accurately communicating with parents about
  individual student performances.
4 = statistical transformations for raw data gathered about
  student performances.
5 = all of the above.

_____ 3. Low grades and marks issued to students will probably:

1 = cause most students to try harder.
2 = affect secondary students more than primary level students.
3 = insure tough academic standards.
4 = carry few negative consequences for most students.

_____ 4. Providing students multiple opportunities or "trials" to
demonstrate attainment of instructional objectives:

1 = is a good idea.
2 = should be used with all students in a course which is
  criterion-referenced.
3 = should be used only with students who have experienced
  difficulty or are in danger of failing.
4 = reduces the reliability and validity of grades.
5 = none of the above.

_____ 5. All other things being equal, which of the following marking
practices will yield the t ,c reliable marks?

1 = assigning numbers one through five.
2 = assigning marks of A, B, C, D, or F.
3 = assigning only pass or fail marks.
4 = using percentile ranks.

_____ 6. Good grades and marks will probably:

1 = increase the motivation level of most students who receive
  them.
2 = have little impact on most students.
3 = affect primary level students more than secondary level
  students.
4 = affect secondary level students more than primary level
  students.

_____ 7. Decisions made in assigning grades and marks:

1 = may be subjected to judicial review.
2 = will not be subjected to judicial review because the courts have traditionally left educational decisions to professional educators.
3 = should be bound by the same due process concerns as decisions to discipline students for misbehavior.
4 = should be reviewed by administrators.

_____ 8. The reliability of grades can be improved by:

1 = including non-cognitive factors in the assessment of student performance.
2 = emphasizing evaluation rather than assessment.
3 = emphasizing gain rather than terminal behavior.
4 = using a two-category system, such as pass-fail or satisfactory-unsatisfactory.
5 = increasing the number of grade categories.

_____ 9. To develop good grading and marking systems, practicing educators should:

1 = adhere to a standardized system throughout an entire school district.
2 = be familiar with systematic measurement techniques.
3 = develop consistent means for scoring student work and assigning grades and marks.
4 = teach and assess the same instructional content in equivalent courses of study.
5 = all of the above.

_____ 10. Grading is defined as the process of:

1 = correcting student work.
2 = assigning cumulative grade symbols to report cards.
3 = attaching a symbol to a student product.
4 = all of the above.
5 = none of the above.

_____ 11. Giving "extra credit" assignments and opportunities for a student to improve a grade:

1 = is a good idea.
2 = should be used only for students who have experienced difficulty competing with classmates in attaining levels of academic achievement.
3 = should be used only with students who are in danger of failing a course.
4 = reduces the reliability and validity of the grade.

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may or may not be congruent with the practices you use).

_____ 12. When combining scores, if a teacher wishes to give equal weight to several scores, best results will be obtained:

1 = if percentile scores are used.
2 = if percentage scores are used.
3 = if raw scores are used.
4 = if standard scores are used.
5 = any of the above—it makes no difference.

_____ 13. Scoring is defined as the process of:

1 = correcting student work and calculating the accuracy of performance.
2 = statistically combining several measures of a student's performance to devise a cumulative score.
3 = calculating the distribution of student performances on a given task.
4 = correcting student work and assigning grade symbols to represent levels of performance.
5 = none of the above.

_____ 14. Reducing grades or marks as a punishment for a student's misbehavior in school:

1 = should be done only in required courses of study.
2 = should only be done as a "last resort" when all other means of discipline have failed.
3 = should generally not be done.
4 = should be done as a matter of routine practice.
5 = none of the above.

_____ 15. If a single grade is given, it should represent:

1 = academic achievement.
2 = attendance.
3 = behavior.
4 = effort.
5 = all of the above.

_____ 16. How do grades determined by level of improvement (growth) compare with grades determined by level of achievement (status)?

1 = Growth grades are more reliable and more valid as measures of proficiency.
2 = Growth grades are more reliable but less valid as measures of proficiency.
3 = Growth grades are less reliable but more valid as measures of proficiency.
4 = Growth grades are less reliable and less valid as measures of proficiency.

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES
AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may
or may not be congruent with the practices you use).

_____ 17. Marks in a course should be assigned and interpreted as:

1 = a combination of absolute measures of achievement, relative
    measures of achievement, and evaluations of students.
2 = measures of the level of student achievement.
3 = measures of growth.
4 = evaluations of students.
5 = none of the above.

_____ 18. Reducing grades as a punishment for tardiness in completing
         an assignment:

1 = is a good idea.
2 = should be done as a matter of policy.
3 = should be used only with those students who are habitually
    late in submitting assignments.
4 = should not be done.

_____ 19. Reducing a student's grade or mark as a sanction for truancy
         should:

1 = be done.
2 = generally not be done.
3 = be done only in required courses of study.
4 = be done only in non-required courses of study.

_____ 20. In a comprehensive system of marking, separate marks should
         be reported for:

1 = academic achievement and attendance.
2 = academic achievement and effort.
3 = academic achievement and behavior.
4 = attendance and effort and achievement.
5 = academic achievement, attendance, effort, and behavior.

_____ 21. Grades and marks should be used to:

1 = communicate the status of student academic achievement.
2 = assist in making decisions about retention or promotion.
3 = assist in determining the eligibility of students for
    participation in the activities curriculum at school.
4 = monitor the academic progress of students.
5 = all of the above.

_____ 22. Which of the following should be considered in designing
         a grading and marking system ?

1 = content validity.
2 = reliability.
3 = "face" validity among teachers.
4 = "face" validity among students and parents.
5 = all of the above.

ITEM WORKSHEET

I. List the numbers of items that represent pieces of information you deem essential for educators to either understand or apply to insure good grading and marking practices.

_____

_____

_____

_____

II. Are there any concepts or issues not included in the test instrument that, in your opinion, should be included? If so, please list them below.

_____

_____

_____

_____

_____

_____

THANK YOU!

811 Ogden
Oxford, Nebraska 68967
(308) 824-3397
November 27, 1984

Three weeks ago, on November 6, 1984, I sent you a letter
asking for your assistance in validating a test on grades
and marks. I have not yet received your response.

If you have already completed the test and the item work-
sheet that were enclosed with my original letter and have
placed them in the mail, please disregard this letter. If
however, you have not yet completed the test and the item
worksheet, would you please consider doing so at this time?
Your responses are extremely important.

I am enclosing duplicate copies of the test and the item
worksheet. Please complete both according to the instructions
which have been provided, and return all of the pages to me
in the enclosed envelope as soon as possible, but not later
than December 10, 1984.

Thank you again for your assistance. If you have any ques-
tions, please feel free to call or write me at your convenience.
My home address and phone number are listed above. My work
phone number is (308) 824-3209.

Sincerely,

Kenneth E. Anderson

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES
AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may
or may not be congruent with the practices you use).

_____ 1. The process of marking:

    1 = assigns some symbol to represent the achievement of a
        student on a single product.
    2 = assigns some symbol to represent the cumulative achieve-
        ment of a student on an articulated set of products.
    3 = assigns a raw or derived score to represent the achieve-
        ment of a student on a given product.
    4 = all of the above.
    5 = none of the above.

_____ 2. Marks are valid as:

    1 = predictors of the future performance of individual students.
    2 = predictors of the future performance of groups of students.
    3 = a means for accurately communicating with parents about
        individual student performances.
    4 = statistical transformations for raw data gathered about
        student performances.
    5 = all of the above.

_____ 3. Low grades and marks issued to students will probably:

    1 = cause most students to try harder.
    2 = affect secondary students more than primary level students.
    3 = insure tough academic standards.
    4 = carry few negative consequences for most students.

_____ 4. Providing students multiple opportunities or "trials" to
demonstrate attainment of instructional objectives:

    1 = is a good idea.
    2 = should be used with all students in a course which is
        criterion-referenced.
    3 = should be used only with students who have experienced
        difficulty or are in danger of failing.
    4 = reduces the reliability and validity of grades.
    5 = none of the above.

_____ 5. All other things being equal, which of the following marking
practices will yield the most reliable marks?

    1 = assigning numbers one through five.
    2 = assigning marks of A, B, C, D, or F.
    3 = assigning only pass or fail marks.
    4 = using percentile ranks.

_____ 6. Good grades and marks will probably:

    1 = increase the motivation level of most students who receive
        them.
    2 = have little impact on most students.
    3 = affect primary level students more than secondary level
        students.
    4 = affect secondary level students more than primary level
        students.

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES
AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may
or may not be contruent with the practices you use).

_____ 7. Decisions made in assigning grades and marks:

1 = may be subjected to judicial review.
2 = will not be subjected to judicial review because the courts
    have traditionally left educational decisions to profes-
    sional educators.
3 = should be bound by the same due process concerns as
    decisions to discipline students for misbehavior.
4 = should be reviewed by administrators.

_____ 8. The reliability of grades can be improved by:

1 = including non-cognitive factors in the assessment of
    student performance.
2 = emphasizing evaluation rather than assessment.
3 = emphasizing gain rather than terminal behavior.
4 = using a two-category system, such as pass-fail or satis-
    factory-unsatisfactory.
5 = increasing the number of grade categories.

_____ 9. To develop good grading and marking systems, practicing
             educators should:

1 = adhere to a standardized system throughout an entire
    school district.
2 = be familiar with systematic measurement techniques.
3 = develop consistent means for scoring student work and
    assigning grades and marks.
4 = teach and assess the same instructional content in
    equivalent courses of study.
5 = all of the above.

_____ 10. Grading is defined as the process of:

1 = correcting student work.
2 = assigning cumulative grade symbols to report cards.
3 = attaching a symbol to a student product.
4 = all of the above.
5 = none of the above.

_____ 11. Giving "extra credit" assignments and opportunities for a
             student to improve a grade:

1 = is a good idea.
2 = should be used only for students who have experienced
    difficulty competing with classmates in attaining levels
    of academic achievement.
3 = should be used only with students who are in danger of
    failing a course.
4 = reduces the reliability and validity of the grade.

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may or may not be congruent with the practices you use).

_____ 12. When combining scores, if a teacher wishes to give equal weight to several scores, best results will be obtained:

1 = if percentile scores are used.
2 = if percentage scores are used.
3 = if raw scores are used.
4 = if standard scores are used.
5 = any of the above—it makes no difference.

_____ 13. Scoring is defined as the process of:

1 = correcting student work and calculating the accuracy of performance.
2 = statistically combining several measures of a student's performance to devise a cumulative score.
3 = calculating the distribution of student performances on a given task.
4 = correcting student work and assigning grade symbols to represent levels of performance.
5 = none of the above.

_____ 14. Reducing grades or marks as a punishment for a student's misbehavior in school:

1 = should be done only in required courses of study.
2 = should only be done as a "last resort" when all other means of discipline have failed.
3 = should generally not be done.
4 = should be done as a matter of routine practice.
5 = none of the above.

_____ 15. If a single grade is given, it should represent:

1 = academic achievement.
2 = attendance.
3 = behavior.
4 = effort.
5 = all of the above.

_____ 16. How do grades determined by level of improvement (growth) compare with grades determined by level of achievement (status)?

1 = Growth grades are more reliable and more valid as measures of proficiency.
2 = Growth grades are more reliable but less valid as measures of proficiency.
3 = Growth grades are less reliable but more valid as measures of proficiency.
4 = Growth grades are less reliable and less valid as measures of proficiency.

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may or may not be congruent with the practices you use).

_____ 17. Marks in a course should be assigned and interpreted as:

    1 = a combination of absolute measures of achievement, relative measures of achievement, and evaluations of students.
    2 = measures of the level of student achievement.
    3 = measures of growth.
    4 = evaluations of students.
    5 = none of the above.

_____ 18. Reducing grades as a punishment for tardiness in completing an assignment:

    1 = is a good idea.
    2 = should be done as a matter of policy.
    3 = should be used only with those students who are habitually late in submitting assignments.
    4 = should not be done.

_____ 19. Reducing a student's grade or mark as a sanction for truancy should:

    1 = be done.
    2 = generally not be done.
    3 = be done only in required courses of study.
    4 = be done only in non-required courses of study.

_____ 20. In a comprehensive system of marking, separate marks should be reported for:

    1 = academic achievement and attendance.
    2 = academic achievement and effort.
    3 = academic achievement and behavior.
    4 = attendance and effort and achievement.
    5 = academic achievement, attendance, effort, and behavior.

_____ 21. Grades and marks should be used to:

    1 = communicate the status of student academic achievement.
    2 = assist in making decisions about retention or promotion.
    3 = assist in determining the eligibility of students for participation in the activities curriculum at school.
    4 = monitor the academic progress of students.
    5 = all of the above.

_____ 22. Which of the following should be considered in designing a grading and marking system ?

    1 = content validity.
    2 = reliability.
    3 = "face" validity among teachers.
    4 = "face" validity among students and parents.
    5 = all of the above.

ITEM WORKSHEET

I. List the numbers of items that represent pieces of information
    you deem essential for educators to either understand or apply
    to insure good grading and marking practices.

    _____

    _____

    _____

    _____

II. Are there any concepts or issues not included in the test in-
    strument that, in your opinion, should be included? If so,
    please list them below.

    _____

    _____

    _____

    _____

    _____

    _____

    _____

THANK YOU!

APPENDIX B

Final Validation Form

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may or may not be congruent with the practices you use).

_____ 1. The process of marking:

   1 = assigns some symbol to represent the achievement of a student on a single product.
   2 = assigns some symbol to represent the cumulative achievement of a student on an articulated set of products.
   3 = assigns a raw or derived score to represent the achievement of a student on a given product.
   4 = all of the above.
   5 = none of the above.

_____ 2. Marks are valid as:

   1 = predictors of the future performance of individual students.
   2 = predictors of the future performance of groups of students.
   3 = a means for accurately communicating with parents about individual student performances.
   4 = statistical transformations for raw data gathered about student performances.
   5 = all of the above.

_____ 3. Low grades and marks issued to students will probably:

   1 = cause most students to try harder.
   2 = affect secondary students more than primary level students.
   3 = insure tough academic standards.
   4 = carry few negative consequences for most students.

_____ 4. Providing students multiple opportunities or "trials" to demonstrate attainment of instructional objectives:

   1 = is a good idea.
   2 = should be used with all students in a course which is criterion-referenced.
   3 = should be used only with students who have experienced difficulty or are in danger of failing.
   4 = reduces the reliability and validity of grades.
   5 = none of the above.

_____ 5. All other things being equal, which of the following marking practices will yield the most reliable marks?

   1 = assigning numbers one through five.
   2 = assigning marks of A, B, C, D, or F.
   3 = assigning only pass or fail marks.
   4 = using percentile ranks.

_____ 6. Good grades and marks will probably:

   1 = increase the motivation level of most students who receive them.
   2 = have little impact on most students.
   3 = affect primary level students more than secondary level students.
   4 = affect secondary level students more than primary level students.

_____ 7. Decisions made in assigning grades and marks:

1 = may be subjected to judicial review.
2 = will not be subjected to judicial review because the courts
    have traditionally left educational decisions to profes-
    sional educators.
3 = should be bound by the same due process concerns as
    decisions to discipline students for misbehavior.
4 = should be reviewed by administrators.

_____ 8. The reliability of grades can be improved by:

1 = including non-cognitive factors in the assessment of
    student performance.
2 = emphasizing evaluation rather than assessment.
3 = emphasizing gain rather than terminal behavior.
4 = using a two-category system, such as pass-fail or satis-
    factory-unsatisfactory.
5 = increasing the number of grade categories.

_____ 9. To develop good grading and marking systems, practicing
educators should:

1 = adhere to a standardized system throughout an entire
    school district.
2 = be familiar with systematic measurement techniques.
3 = develop consistent means for scoring student work and
    assigning grades and marks.
4 = teach and assess the same instructional content in
    equivalent courses of study.
5 = all of the above.

_____ 10. Grading is defined as the process of:

1 = correcting student work.
2 = assigning cumulative grade symbols to report cards.
3 = attaching a symbol to a student product.
4 = all of the above.
5 = none of the above.

_____ 11. Giving "extra credit" assignments and opportunities for a
student to improve a grade:

1 = is a good idea.
2 = should be used only for students who have experienced
    difficulty competing with classmates in attaining levels
    of academic achievement.
3 = should be used only with students who are in danger of
    failing a course.
4 = reduces the reliability and validity of the grade.

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES
AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may
or may not be congruent with the practices you use).

_____ 12. When combining scores, if a teacher wishes to give equal
weight to several scores, best results will be obtained:

1 = if percentile scores are used.
2 = if percentage scores are used.
3 = if raw scores are used.
4 = if standard scores are used.
5 = any of the above—it makes no difference.

_____ 13. Scoring is defined as the process of:

1 = correcting student work and calculating the accuracy of
performance.
2 = statistically combining several measures of a student's
performance to devise a cumulative score.
3 = calculating the distribution of student performances on
a given task.
4 = correcting student work and assigning grade symbols to
represent levels of performance.
5 = none of the above.

_____ 14. Reducing grades or marks as a punishment for a student's
misbehavior in school:

1 = should be done only in required courses of study.
2 = should only be done as a "last resort" when all other
means of discipline have failed.
3 = should generally not be done.
4 = should be done as a matter of routine practice.
5 = none of the above.

_____ 15. If a single grade is given, it should represent:

1 = academic achievement.
2 = attendance.
3 = behavior.
4 = effort.
5 = all of the above.

_____ 16. How do grades determined by level of improvement (growth) com-
pare with grades determined by level of achievement (status)?

1 = Growth grades are more reliable and more valid as measures
of proficiency.
2 = Growth grades are more reliable but less valid as mea-
sures of proficiency.
3 = Growth grades are less reliable but more valid as mea-
sures of proficiency.
4 = Growth grades are less reliable and less valid as mea-
sures of proficiency.

SELECT THE ANSWER WHICH BEST DESCRIBES EITHER THE ATTRIBUTES OF GRADES AND MARKS OR RECOMMENDED GRADING AND MARKING PRACTICES (this answer may or may not be congruent with the practices you use).

_____ 17. Marks in a course should be assigned and interpreted as:

1 = a combination of absolute measures of achievement, relative measures of achievement, and evaluations of students.
2 = measures of the level of student achievement.
3 = measures of growth.
4 = evaluations of students.
5 = none of the above.

_____ 18. Reducing grades as a punishment for tardiness in completing an assignment:

1 = is a good idea.
2 = should be done as a matter of policy.
3 = should be used only with those students who are habitually late in submitting assignments.
4 = should not be done.

_____ 19. Reducing a student's grade or mark as a sanction for truancy should:

1 = be done.
2 = generally not be done.
3 = be done only in required courses of study.
4 = be done only in non-required courses of study.

_____ 20. In a comprehensive system of marking, separate marks should be reported for:

1 = academic achievement and attendance.
2 = academic achievement and effort.
3 = academic achievement and behavior.
4 = attendance and effort and achievement.
5 = academic achievement, attendance, effort, and behavior.

_____ 21. Grades and marks should be used to:

1 = communicate the status of student academic achievement.
2 = assist in making decisions about retention or promotion.
3 = assist in determining the eligibility of students for participation in the activities curriculum at school.
4 = monitor the academic progress of students.
5 = all of the above.

_____ 22. Which of the following should be considered in designing a grading and marking system ?

1 = content validity.
2 = reliability.
3 = "face" validity among teachers.
4 = "face" validity among students and parents.
5 = all of the above.